Profiliranje spletnih uporabnikov v spletnem oglaševanju

Domen Košir

doktorska disertacija

predana

Fakulteti za računalništvo in informatiko

kot del izpolnjevanja pogojev za pridobitev naziva

doktor znanosti

s področja

računalništva in informatike

Ljubljana, 2015

Profiliranje spletnih uporabnikov v spletnem oglaševanju

Domen Košir

doktorska disertacija

predana

Fakulteti za računalništvo in informatiko

kot del izpolnjevanja pogojev za pridobitev naziva

doktor znanosti

s področja

računalništva in informatike

Ljubljana, 2015

IZJAVA

Izjavljam, da sem avtor dela in da slednje ne vsebuje materiala, ki bi ga kdorkoli predhodno že objavil ali oddal v obravnavo za pridobitev naziva na univerzi ali na drugem visokošolskem zavodu, razen v primerih, kjer so navedeni viri.

— Domen Košir —

februar 2015

Oddajo so odobrili

dr. Igor Kononenko

redni profesor za računalništvo in informatiko

MENTOR IN ČLAN OCENJEVALNE KOMISIJE

dr. Zoran Bosnić

izredni profesor za računalništvo in informatiko

MENTOR IN ČLAN OCENJEVALNE KOMISIJE

dr. Franc Solina

redni profesor za računalništvo in informatiko

PREDSEDNIK OCENJEVALNE KOMISIJE

dr. Dunja Mladenić

redna profesorica za računalništvo

ZUNANJA ČLANICA OCENJEVALNE KOMISIJE

Institut Jožef Stefan

PREDHODNA OBJAVA

Izjavljam, da so bili rezultati obravnavane raziskave predhodno objavljeni/sprejeti za objavo v recenzirani reviji ali javno predstavljeni v naslednjih primerih:

[1] D. Košir, I. Kononenko, and Z. Bosnić. Web User Profiles with Time-Decay and Prototyping. *Applied Intelligence*, 233(2):199–220, 2014.
doi: 10.1016/j.jtbi.2004.10.003

Potrjujem, da sem pridobil pisna dovoljenja vseh lastnikov avtorskih pravic, ki mi dovoljujejo vključitev zgoraj navedenega materiala v pričujočo disertacijo. Potrjujem, da zgoraj navedeni material opisuje rezultate raziskav, izvedenih v času mojega podiplomskega študija na Univerzi v Ljubljani.

Moji Katji.

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*
Domen Košir
Profiliranje spletnih uporabnikov v spletnem oglaševanju

POVZETEK

Dobički velikih spletnih podjetij, ki se dandanes merijo v milijardah dolarjev, izhajajo večinoma z naslova spletnega oglaševanja in so glavno gonilo napredka na področjih profiliranja spletnih uporabnikov ter sistemov za priporočanje. Že majhne spremembe v kvaliteti priporočil imajo lahko velike ekonomske posledice, zato se oglaševalska podjetja nenehno trudijo izboljšati obstoječe metode za profiliranje spletnih uporabnikov in njihovo rabo v oglaševanju.

V disertaciji je predstavljena nova metoda za gradnjo ontoloških profilov spletnih uporabnikov AverageActionFC, ki uporablja tehniki časovnega pozabljanja in popravljanja profilov s prototipi. S časovnim pozabljanjem dodeljujemo pomembnost posameznim preteklim dogodkom iz uporabnikovega klikotoka, s prototipi pa obogatimo njegov profil z domenskim znanjem in tako izboljšamo kvaliteto profila. S poskusi na dveh velikih in realnih podatkovnih množicah pokažemo, kako se giblje kvaliteta uporabnikovega profila glede na njegovo starost in količino podatkov, ki smo jih uporabili pri profiliranju. Rezultati kažejo, da lahko z našo metodo zgradimo profile višje kakovosti kot z obstoječimi metodami.

Uporabniške profile uporabljamo v sistemih za priporočanje z izbiranjem na podlagi sodelovanja, ki temeljijo na matrični faktorizaciji, za premagovanje t.i. hladnega zagona. Vrednosti skritih faktorjev za nove uporabnike napovedujemo na podlagi semantičnih informacij v profilih z uporabo metod strojnega učenja in tako izboljšamo kakovost seznamov priporočil. Kakovost teh seznamov izdatno izboljšamo s pametnim kombiniranjem priporočil več sistemov za priporočanje.

Področji profiliranja spletnih uporabnikov in priporočilnih sistemov sta relativno mladi in aktivni. V nadaljnjem delu se bomo posvetili izboljšavam razvitih metod, pri čemer imamo v mislih predvsem iterativno gradnjo profilov in priporočanje na podlagi vedenja o popularnosti posameznih tematik ob različnih časih. Sisteme za priporočanje bomo poskusili dodatno izboljšati še z upoštevanjem negativnih povratnih informacij spletnih uporabnikov.

Ključne besede: spletno oglaševanje, profiliranje, sistemi za priporočanje, sledenje uporabnikom, zasebnost uporabnikov

University *of Ljubljana*
Faculty *of Computer and Information Science*
Domen Košir
Web User Profiling in Online Advertising

ABSTRACT

Online advertising is a multi-billion dollar industry. Most of the profits of large internet companies come from online advertising and even small improvements of profiling methods can give a company an edge against its competitors. This can only be achieved through better understanding of web user behavior and applying this knowledge to improve user profiling methods and recommendation systems.

In this thesis, we present a novel ontological profiling method AverageActionFC. It uses time-based forgetting to assign importance to past events in the user's clickstream and profile correction with prototypes to improve the quality of the user's profile with domain knowledge. The experiments on two large, real-world datasets were designed to reveal how the age of the profile and the length of the learning clickstream affect the quality of the user's profile. The results show that our method significantly outperforms existing methods.

Collaborative filtering recommendation systems suffer from the cold start problem. We tackle this problem by using the semantic information in the user profiles to improve the quality of recommendations from matrix factorization-based systems. We employ machine learning algorithms to build models which use the values of semantic attributes in the profiles to predict the values of the latent factors for new users. Based on observations of the quality of recommendations from different systems, we further improve the quality of recommendation lists by combining recommendations from two or more systems.

User profiling and recommendation systems are relatively young and active research fields. In the future, we will adapt our profiling method for the web environment by using iterative techniques to build the user interest models. We will focus on utilizing the knowledge of the popularity of different topics over time to produce better recommendations to web users. The quality of recommendations can also be improved by gathering and taking into account negative user feedback.

Key words: online advertising, profiling, recommendation systems, user tracking, privacy

ZAHVALA

Po vseh presedenih nočeh in prespanih jutrih me ob zaključku doktorskega študija navdajajo mešani občutki, saj se končuje dokaj stresno, a kljub temu prijetno in uspešno obdobje v mojem življenju. A vsega (lepega) je enkrat konec in omeniti moram vse, ki so pri tem igrali pomembno vlogo.

Najprej bi se rad zahvalil svojima mentorjema prof. dr. Igorju Kononenku in izr. prof. dr. Zoranu Bosniću, ki sta me usmerjala in mi pomagala pri raziskovalnem delu. Pot do cilja sta mi olajšala ne le s strokovnostjo, temveč tudi s prijetnimi odnosi in nalezljivim optimizmom.

Zahvalil bi se rad še prof. dr. Dunji Mladenić in prof. dr. Francu Solini, ki sta kot člana komisije za oceno doktorske disertacije s konstruktivnimi komentarji pomembno prispevala h kvaliteti te disertacije.

Hvala tudi vodstvu in vsem zaposlenim pri podjetju Httpool d.o.o., ki so mi omogočili nadaljevanje študija, poigravanje s podatki in nudili pomoč pri birokraciji ter raznoraznih zapletih na strežniku za obdelavo podatkov.

Hvaležen sem tudi ostalim članom Laboratorija za kognitivno modeliranje ter zaposlenim na Fakulteti za računalništvo in informatiko, ki so mi kot izjemen kolektiv pomagali pri marsikateri težavi.

Mojima staršema Valeriji in Janku bi se še posebej zahvalil za dolga leta podpore in ljubezni, bratu Janu, razširjeni družini in prijateljem pa za dragocene trenutke sprostitve.

Na koncu bi se rad zahvalil še moji bodoči ženi Katji za vso ljubezen, razumevanje, pomoč in dolge sprehode po Barju, ki sem jih v stresnih trenutkih res potreboval.

— Domen Košir, Ljubljana, februar 2015.

KAZALO

I

Uvod

1.1 Oglaševanje

Oglaševanje je ena od oblik tržnega komuniciranja. Ima dva glavna namena:

1. informiranje in prepričevanje potrošnikov k nakupu določenega izdelka oz. storitve ali
2. večanje prepoznavnosti blagovne znamke ali podjetja.

Definicija oglaševanja se je spreminjala s časom. Predvsem je bila odvisna od medijev in načinov komunikacije, ki so jih mediji omogočali. V zadnjem času je popularna naslednja definicija [1, 2]:

Definicija 1: Oglaševanje je plačana, skozi medij posredovana oblika komunikacije prepoznavnega izvora, oblikovana, da prepriča prejemnika, da nekaj stori, bodisi takoj bodisi v prihodnosti.

Jančič [2] predlaga še alternativno definicijo, ki poudarja, da lahko v poplavi vsebin le kreativna komunikacija doseže svoj učinek. Z izrazom "izpolnljiva obljuba" distancira oglaševalski nagovor, katerega bistvo je informiranje in prepričevanje uporabnika z dajanjem stvarnih in uresničljivih obljub, od reklamnega in propagandnega, v okviru katerih lahko uporabnika zavajamo in ga nagovarjamo z lažnimi oz. praznimi obljubami. V njegovi definiciji oglaševanje ni nujno plačana komunikacija, saj se v zadnjem času pojavljajo tudi t.i. brezplačni oglasi, pri katerih so naročniki pogosto dobrodelne organizacije.

Definicija 2: Oglaševanje je načrtovana, naročena in podpisana kreativna (množična) komunikacija, katere namen je spodbujanje procesov menjave med ponudniki in porabniki s podajanjem izpolnljivih obljub.

Na področju oglaševanja sicer vlada terminološka zmeda. Poleg izraza oglaševanje se pojavljajo še izrazi, kot so reklamiranje, propaganda, promocija itd. Jančič [2] razlikuje med prej naštetimi izrazi z razlikovanjem med uporabljenimi pristopi in vsebinami:

- *Reklama* lahko vključuje vsiljive in zavajajoče trditve, velikokrat pretirava v hvaljenju in opisu lastnostih izdelkov oz. storitev, ki jih skuša prodati.
- *Propaganda* razširja nauke in prepričanja, glede ponujanih izdelkov oz. storitev pa pogosto podaja nejasne obljube, brez konkretnih ciljev ali rokov.
- *Promocija* je splošen izraz, ki ponavadi zaobjema oglaševanje, publiciteto, osebno prodajo, pospeševanje prodaje in neposredno prodajo.

1.1.1 Spletno oglaševanje

Internet je zaradi velike popularnosti in globalne razširjenosti hitro postal zelo zanimiv medij za oglaševalsko industrijo.

V primerjavi z oglaševanjem na tradicionalnih medijih ima spletno oglaševanje naslednje prednosti:

- *Hitra in nezahtevna izdelava oglasa* - Spletni oglaševalski servisi navadno ponujajo oglaševalcem oz. agencijam spletne vmesnike, s katerimi je možno oglas izdelati v nekaj minutah. Izdelani oglasi se lahko avtomatično barvno in stilsko prilagajajo spletnim stranem, na katerih se prikazujejo.
- Oglaševanje je *fleksibilno*, saj lahko oglase kadarkoli spremenimo in prilagodimo ciljni publiki.
- *Takojšnji globalni dostop* - Izdelan oglas je takoj pripravljen za prikazovanje spletnim uporabnikom ne glede na njihovo fizično lokacijo.
- *Ciljano oglaševanje* - Sodobna tehnologija nam omogoča personalizirano prikazovanje oglasov. To pomeni, da lahko vsakemu posameznemu uporabniku spleta prikažemo oglas, ki je zanj najbolj primeren.
- *Merjenje uspešnosti oglaševanja* - S preprostim sledenjem klikom na oglase in nakupom uporabnikov lahko zelo zanesljivo ugotovimo, kako uspešna je posamezna oglaševalska akcija.
- V zadnjem času ima vedno večji pomen *interaktivnost*, ki temelji na dvosmerni komunikaciji med oglaševalcem in publiko, ter je najbolj izražena pri oglaševanju na socialnih omrežjih in v nagradnih igrah.

Med negativnimi učinki spletnega oglaševanja se največkrat omenja *zasičenost* tega medija z oglasi. Ta je postala še posebej očitna v zadnjih letih, ko so se v večini medijskih hiš zaradi upada naklad tiskanih revij in časopisov začeli bolj posvečati spletnim medijem.

Na spletu so se tekom let razvile različne oblike oglaševanja:

- *Elektronska pošta* je eden najstarejših še uporabljanih komunikacijskih kanalov na internetu. Problem nezaželene e-pošte (angl. spam) je še vedno aktualen. Ocenjuje se, da zavzema nezaželena okoli dve tretjini vse poslane e-pošte.
- Na spletnih straneh se v okviru t.i. *prikaznega spletnega oglaševanja* (angl. online display advertising) obiskovalcem na spletnih straneh prikazujejo pasice, tekstovni, slikovni in video oglasi. Oglasi se ponavadi pojavljajo v posebej za oglase namenjenih delih spletnih strani ali v pojavnih (angl. pop-up) oknih.
- V *spletnih iskalnikih* se lahko uporabniku poleg rezultatov iskanja prikažejo še oglasi. Ta vrsta oglaševanja se je izkazala za zelo uspešno, saj je med iskanjem informacij uporabnik zelo dovzeten za oglasna sporočila, ki so v skladu z njegovimi poizvedbami.
- *Oglaševanje v socialnih omrežjih* lahko učinkovito izkorišča informacije o uporabnikovih prijateljih in interesih, ki jih je uporabnik sam vnesel v svoj profil.
- Oglasom smo pogosto izpostavljeni, ko nam je nek izdelek ali storitev na voljo brez plačila, npr. *v nagradnih igrah* in pri uporabi *brezplačnih mobilnih aplikacij*.

S sledenjem obiskom spletnega uporabnika in analiziranjem obiskanih vsebin lahko zgradimo model njegovih interesov, kar zelo pripomore k učinkovitosti vseh prej naštetih tipov spletnega oglaševanja.

1.2 Motivacija

Vedno bolj očitni selitvi novic in ostalih vsebin iz tiskanih medijev na splet zvesto sledi tudi oglaševanje. S povečevanjem denarnih vložkov oglaševalcev in konkurence na spletu postajajo napredne oblike oglaševanja nujen del ponudbe ponudnikov oglaševanja.

Za ciljano oglaševanje je potrebno vsaj do neke poznati spletno populacijo, da lahko posameznim uporabnikom prikazujemo za njih čim bolj zanimive oglase. Za zbiranje

podatkov o spletnih uporabnikih so se v začetku zelo pogosto uporabljali vprašalniki in ankete. Taki pristopi, ki zahtevajo od uporabnika aktivno sodelovanje, so vedno manj uspešni, saj za razliko od vse bolj popularnih nagradnih iger tu uporabnik nima od sodelovanja nikakršne koristi. V zadnjih letih se največkrat uporabljajo neinvazivni pristopi, ki temeljijo na spremljanju uporabnikovih aktivnosti na spletu, pri čemer si ponudniki oglaševanja največkrat pomagajo s sledenjem uporabnikom s pomočjo spletnih piškotkov. Taki pristopi so bili dolgo časa povsem sprejemljivi tako za uporabnike kot za ponudnike oglaševanja. Veljal je nekakšen *status quo*. Vedno večje zavedanje uporabnikov o pomembnosti varovanja zasebnosti na spletu je povzročilo spremembe zakonodaje, ki od leta 2013 [3] v Sloveniji omejuje uporabo piškotkov za sledenje spletnim uporabnikom. Novi Zakon o elektronskih komunikacijah ZEKom-1 dovoljuje uporabo sledilnih piškotkov na spletnih straneh le z izrecnim dovoljenjem uporabnika. Zanesljivost sledenja spletnim uporabnikom je v zadnjem času upadla, saj večina modernih brskalnikov podpira tudi t.i. "zasebni način" delovanja, ki onemogoča trajno hrambo podatkov na uporabnikovem računalniku, vključno s hrambo morebitnega uporabnikovega dovoljenja za sledenje.

Podobna zakonodaja velja tudi drugod v EU, kar je nedvomno prisililo mnoga podjetja v razmislek in razvoj alternativ. Poleg spletnih piškotkov obstajajo tudi drugi, zanimivi in dostikrat sporni načini za sledenje spletnim uporabnikom, ki bi se jih dalo uporabiti za potrebe spletnega oglaševanja. Med seboj se razlikujejo v pristopu, zanesljivosti in zakonitosti uporabe. Borba proti nekaterim tehnnikam zahteva od spletnega uporabnika kar nekaj znanja in energije, pri kombinacijah različnih sledilnih tehnik pa je anonimnost na internetu za običajne spletne uporabnike v praksi nedosegljiva. Podatki o spletnih uporabnikih, ki jih na tak ali drugačen način zbirajo različna podjetja in posamezniki, se na koncu največkrat uporabijo za potrebe oglaševanja.

V okviru te disertacije je predstavljena nova metoda za gradnjo ontoloških profilov spletnih uporabnikov AverageActionFC. Obstoječe metode uporabljajo za gradnjo uporabnikovega profila le informacije o njegovih lastnih preteklih aktivnostih, ki pa imajo, še posebej v primeru, da o uporabniku ne vemo veliko, zelo majhno informativno vrednost. Nova metoda obogati uporabnikov profil še z informacijami o interesih njemu podobnih uporabnikov, kar lahko pomembno prispeva k zvišanju kakovosti zgrajenega profila. Velika večina raziskav na področju profiliranja spletnih uporabnikov je bila opravljena na majhnih ali umetnih podatkovnih množicah, kar zmanjšuje zaupanje v objavljene rezultate. S poskusi na obsežnih in realnih podatkih smo izvedli

verodostojno primerjavo med metodami za profiliranje uporabnikov.

Uporaba profilov spletnih uporabnikov v spletnih aplikacijah se lahko zdi preprosta in včasih tudi je. Preprosti sistemi za priporočanje na podlagi vsebinskega izbiranja temeljijo le na podobnosti med profilom uporabnika in vsebino posameznega izdelka ali storitve, ki jo želimo priporočiti uporabniku. Z boljšimi uporabniškimi profili lahko dajemo uporabnikom boljša priporočila, s tem pa pomembno pripomoremo h komercialnemu uspehu podjetja, ki to prodaja. Številne raziskave zadnjih let kažejo, da dajejo sistemi za priporočanje na osnovi sodelovanja, ki delujejo na osnovi matrične faktorizacije, dosti boljše predloge, trpijo pa med drugim za t.i. problemom hladnega zagona, do katerega pride, ko pride s sistemom v stik nov uporabnik, o katerem ni znanega dovolj, da bi mu sistem lahko dal kvalitetne predloge. V okviru svojih raziskav smo se lotili reševanja tega problema z uporabo semantičnih informacij v uporabniških profilih.

1.3 Prispevki k znanosti

V tej disertaciji so predstavljeni naslednji prispevki k znanosti:

- *Nova metoda za profiliranje spletnih uporabnikov AverageActionFC*: Za razliko od obstoječih metod, ki gradijo uporabnikov profil le na podlagi njegovih lastnih preteklih aktivnosti, metoda AverageActionFC obogati uporabnikov profil še z domenskim znanjem, ki je prestavljeno v obliki prototipnih profilov. Množico prototipnih profilov lahko definiramo z analizo preteklih aktivnosti celotne uporabniške populacije, v primeru pomanjkanja podatkov pa s pomočjo domenskih strokovnjakov. Metodo AverageActionFC smo primerjali z obstoječimi metodami na dveh realnih podatkovnih množicah: podatki spletne oglaševalske mreže in spletne učilnice.

- *Nov pristop za reševanje problema hladnega zagona v sistemih za priporočanje na osnovi matrične faktorizacije z učenjem iz uporabniških profilov*: O problemu hladnega zagona govorimo, ko pride v stik s sistemom za priporočanje nov uporabnik, o katerem nimamo dovolj podatkov, da bi mu lahko podali kvalitetna priporočila. Za reševanje tega problema smo uporabili semantične informacije v uporabniških profilih in s pomočjo algoritmov strojnega učenja zgradili napovedni model, ki na podlagi semantičnih atributov napoveduje skrite faktorje, ki se uporabljajo za gradnjo seznama priporočil.

- *Pametno kombiniranje priporočil sistemov za priporočanje*: Poskusi na podatkih spletne oglaševalske mreže so pokazali, da dajejo sistemi za priporočanje z izbiranjem s sodelovanjem le nekaj zelo dobrih predlogov, za ostala priporočila pa je bolje uporabiti vsebinsko izbiranje. S pametnim združevanjem priporočil posameznih sistemov za priporočanje smo izdelali kombiniran seznam priporočil, ki se je izkazal za občutno boljšega.

1.4 Pregled disertacije

Rdeča nit disertacije je profiliranje spletnih uporabnikov za potrebe spletnega oglaševanja.

Za učinkovito spletno oglaševaje je nujno potreben dostop do informacij o spletnih uporabnikih in njihovih preteklih aktivnostih. Drugo poglavje vsebuje pregled različnih tehnik za sledenje spletnim uporabnikom ter pregled področja profiliranja spletnih uporabnikov.

Tretje poglavje vsebuje opis nove metode za gradnjo ontoloških profilov spletnih uporabnikov, imenovane AverageActionFC. Metodo primerjamo z nekaterimi obstoječimi metodami za profiliranje na dveh realnih podatkovnih množicah. Pri gradnji profilov uporabnikov spletne oglaševalske mreže Httpool in študentov v spletni učilnici smo z uporabo naše metode bistveno izboljšali kvaliteto profilov v primerjavi z obstoječimi metodami. Metodologija testiranja je zasnovana tako, da omogoča vpogled v gibanje kvalitete uporabniških profilov glede na njihovo starost in dolžino uporabnikove zgodovine, ki je bila uporabljena za gradnjo profilov.

V četrtem poglavju je predstavljena uporaba uporabniških profilov v sistemih za priporočanje. Med seboj so primerjani sistemi, ki delujejo z vsebinskim izbiranjem, izbiranjem na podlagi sodelovanja in hibridni sistemi za priporočanje. V prvi skupini poskusov smo izboljšali sistem za priporočanje na podlagi sodelovanja tako, da smo izkoristili informacije v uporabniških profilih za premagovanje t.i. *problema hladnega zagona*. V drugi skupini poskusov smo s kombiniranjem priporočil več različnih sistemov uspeli še dodatno izboljšati performančne metrike.

V petem poglavju so predstavljeni zaključki in nekatere možne nadaljnje smeri raziskovanja, ki jih na tako svežem in aktualnem področju, kot je spletno oglaševanje, ni malo.

K disertaciji so priloženi še trije dodatki, v katerih ponujamo širši pogled na problematiko zasebnosti, vključno z ustavnopravno ter zakonsko zaščito spletnih uporab-

nikov v Sloveniji, analizo učinkovitosti oglasov glede na njihovo umestitev v spletne strani ter opis gradnje uporabniških profilov z uporabo sinusne regresije.

2

Sledenje in profiliranje spletnih uporabnikov

2.1 Sledenje spletnim uporabnikom

Za oglaševalski sistem je zelo pomembno, da lahko vsak zahtevek za oglas, ki pride do njega, poveže s svojo zbirko podatkov o uporabnikih. Le na ta način lahko spletnemu uporabniku prikaže oglase, ki so zanj zanimivi, in tako poveča verjetnost klika na oglas.

Pri povezovanju zahtevka z uporabniškim profilom poznamo dva pristopa:

- identifikacija uporabnika s številko ID in
- razločevanje med posameznimi uporabniki na podlagi razlik med njimi.

2.1.1 Sledenje z identifikacijo uporabnikov

Pri tem pristopu je bistveno, da na uporabnikovo napravo shranimo identifikator uporabnika (ID). Ko uporabnik pride prvič v stik z oglaševalskim sistemom, mu sistem dodeli številko ID (to je ponavadi zaporedna številka ali naključno zaporedje znakov) in jo shrani nekam na uporabnikov računalnik. Pri vsaki naslednji interakciji z oglaševalskim sistemom identificiramo uporabnika s številko ID, ki jo preberemo z njegovega računalnika.

Učinkovitost in zanesljivost tega pristopa je zelo odvisna od lokacije na uporabnikovem računalniku, kamor shranjujemo uporabnikovo številko ID. V nadaljevanju si bomo podrobneje pogledali nekatere najpogosteje uporabljene lokacije.

Piškotki HTTP

Piškotki HTTP (tudi spletni piškotki) so majhne datoteke, ki jih ustvarja spletni brskalnik. Namenjene so vzdrževanju sej in shranjevanju majhne količine podakov (do 4 KB). Uporaba vsakega piškotka je ponavadi omejena na določeno spletno domeno ali poddomeno, veljavnost pa je omejena s časom trajanja. Sledenje uporabnikom s piškotki je zelo pogosto, a vedno manj zanesljivo zaradi enostavnosti brisanja piškotkov in uporabe zasebnostnega načina delovanja brskalnika.

Piškotki Flash

Za uporabo t.i. piškotkov Flash mora biti v uporabnikovem brskalniku nameščen vtičnik Flash. Vtičnik Flash je zelo razširjen - ocenjuje se, da je nameščen na več kot 90% računalnikih, ki so povezani na svetovni splet. Piškotki Flash delujejo na podoben način kot piškotki HTTP, za razliko od slednjih pa imajo neomejeno trajanje in veliko

večjo kapaciteto (do 100 KB podatkov). Z vidika sledenja uporabnikom imajo piškotki Flash še eno veliko prednost - omogočajo namreč sledenje uporabniku ne glede na to, kateri brskalnik trenutno uporablja - piškotki Flash se namreč shranjujejo v uporabnikovo domačo mapo na računalniku. Na podoben način se da na uporabnikov računalnik shranjevati podatke tudi z Microsoftovim vtičnikom Silverlight.

Skladišče HTML5

Tehnologija HTML5 daje razvijalcem spletnih strani na voljo dve skladišči za podatke. Sejno skladišče (angl. session storage) je uporabno predvsem za shranjevanje začasnih sejnih podatkov, saj obstaja le za čas trenutne seje. Podatki v lokalnem skladišču (angl. local storage) pa so na voljo vse dokler jih uporabnik ali spletna stran ne izbrišeta. Kapaciteta obeh skladišč več kot zadošča potrebam sledenja uporabniku, saj lahko skladišči vsebujeta nekaj MB podatkov na spletno domeno. Za sledenje uporabnikom je zaradi večje trajnosti uporabno predvsem lokalno skladišče.

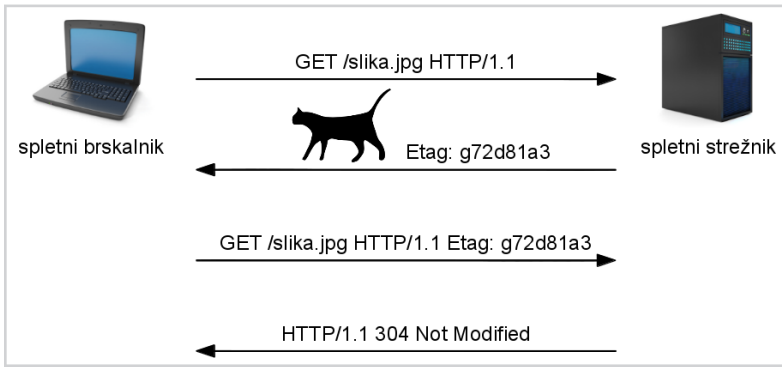
Brisanje vsebin iz obeh skladišč HTML5 lahko predstavlja problem. Nekateri brskalniki namreč uporabniku ne nudijo opcije brisanja vsebin iz HTML5 skladišč - to je možno le preko tonamenskih vtičnikov.

Začasni pomnilnik brskalnika

Ob vsakem obisku spletne strani se vzpostavi komunikacija med uporabnikovim brskalnikom in spletnim strežnikom. Vsebina obiskane spletne strani je lahko sestavljena iz več datotek (datoteke HTML, JavaScript, slike ipd.), vsaki izmed njih pa lahko spletni strežnik pripne še meta informacije, med katerimi je za sledenje spletnim uporabnikom najpomembnejša oznaka ETag, ki sicer služi validaciji vsebin datotek v brskalnikovem začasnem pomnilniku.

Sledenje z uporabo oznake ETag (slika 2.1) poteka tako, da ji spletni strežnik namesto kontrolne vsote (angl. checksum) pripiše unikaten identifikator. Ko uporabnik naslednjič obišče spletno stran, pošlje brskalnik spletnemu strežniku ETag oznake vseh datotek iz začanega pomnilnika, ki pripadajo spletni strani. Namen pošiljanja ETag oznak je večja učinkovitost pri porabi pasovne širine, na ta način pa se jih izkoristi tudi za sledenje uporabnikom.

Temu načinu sledenja se lahko v veliki meri izognemo z rednim brisanjem vsebine začasnega pomnilnika ali uporabo zasebnostnega načina delovanja spletnega brskalnika. Ob vklopu tega načina se namreč spletni dokumenti HTML in pomožne datoteke



Slika 2.1

Prikaz komunikacije med spletnim brskalnikom in spletnim strežnikom. Ob prvem obisku spletne strani zahtevek HTTP spletnega brskalnika za sliko "slika.jpg" ne vsebuje oznake ETag, strežnik pa pošlje poleg datoteke še njeno oznako ETag. Ob vsakem naslednjem obisku pošlje brskalnik v zahtevku tudi oznako ETag, ki lahko služi kot identifikator uporabnika.

(slike, datoteke JavaScript ipd.) shranjujejo v začasni pomnilnik brskalnika le za čas seje.

Evercookie

Prej našteta mesta za shranjevanje uporabnikove številke ID še zdaleč niso edina možna. V odprtokodni knjižnici evercookie [4], ki vsebuje implementacije različnih tehnik za shranjevanje in obnavljanje uporabnikove številke ID, se med drugim uporabljajo tudi:

- zgodovina obiskanih strani (angl. browser history), ki je zapisana v brskalnikovih datotekah,
- brskalnikov začasni spomin (angl. browser cache),
- po meri narejene slike PNG, ki jih vstavimo v spletne strani ter
- razni hrošči in luknje v nameščeni programski opremi.

Knjižnica evercookie uporablja za shranjevanje številke ID kar 13 mest na uporabnikovem računalniku, ob vsakem brisanju podatkov pa jih tudi obnavlja. Spletni uporabniki se tako samo z brisanjem piškotkov in začasnega pomnilnika brskalnika ne morejo ubraniti sledenju.

2.1.2 *Sledenje z razločevanjem uporabnikov*

Spletne uporabnike se da uspešno razločevati med seboj tudi brez shranjevanja podatkov na njihove računalnike. Z analizo zahtevkov, ki prihajajo iz uporabnikovega

računalnika, se da velikokrat zelo natančno ugotoviti, s katerega računalnika je zahtevek prišel. Tak način sledenja dostikrat označujemo z izrazom "fingerprinting". Za sledilca je tak pristop veliko bolj zahteven od shranjevanja in branja številke ID, za uporabnika pa je največkrat neopazen.

Naslov IP

Vsak zahtevek za oglas, ki pride do oglaševalskega sistema, nosi informacijo o izvoru zahtevka. V nasprotju z mnenjem mnogih pa je naslov IP sam po sebi praktično neuporaben podatek. Že v vsakem stanovanju imamo namreč navadno več naprav (namizni, prenosni in tablični računalniki, pametni telefoni ipd.), s katerimi je možno dostopati do spleta, zahtevki z vseh pa bodo imeli isti naslov IP. Problem postane še veliko večji, če upoštevamo, da ljudje veliko brskajo po spletu v podjetjih, šolah in drugih ustanovah, svoj naslov pa lahko skrijejo tudi z uporabo medstrežnikov (angl. proxy server).

Podatki v zahtevku HTTP

Vsak zahtevek HTTP vsebuje polja, ki vsebujejo podatke o uporabnikovem računalniku, operacijskem sistemu, brskalniku in nastavitvah, kot so uporabnikove jezikovne preference in tipi kompresij, ki jih brskalnik podpira. Kombinacija vseh naštetih podatkov lahko že zelo natančno definira računalnik, iz katerega je prišel zahtevek. Vsakega od naštetih podatkov lahko uporabnik brez večjih problemov spremeni - obstajajo tudi vtičniki, ki to olajšajo - vendar pa se večina uporabnikov tega ne poslužuje.

Platno HTML5

Vsi moderni brskalniki podpirajo tehnologiji HTML5 in JavaScript. Z uporabo skripte JavaScript lahko na spletni strani ustvarimo platno, ga uporabniku skrijemo in nanj narišemo nekaj črt ter napisov v različnih pisavah. Naprave spletnih uporabnikov imajo različne strojne (enote CPU, grafični procesorji) in programske karakteristike (različni algoritmi za senčenje), zato prihaja pri izrisovanju na platno do majhnih razlik med slikami, izdelanimi na različnih napravah. Tako ustvarjeno sliko, ki je uporabnik niti ne vidi, lahko uporabimo kot prstni odtis in jo uporabimo za identifikacijo računalnika. V poskusih, ki smo jih izvedli na nekaj več kot 30 napravah, se je ta pristop izkazal za izjemno natančnega, saj je pri namiznih računalnikih za razlikovanje med dvema povsem enakima računalnikoma dovolj že druga različica uporabljenega brskalnika.

Razlikovati nam ni uspelo le med dvema povsem enakima mobilnima telefonoma z enakima brskalnikoma.

Nameščeni vtičniki in pisave

Spletni brskalniki, v katerih je omogočena uporaba jezika JavaScript, nudijo razvijalcem spletnih strani dostop do seznama vseh imen in različic vtičnikov, ki so nameščeni v brskalniku. Preko vtičnikov Flash ali Java lahko pridemo tudi do seznama vseh nameščenih pisav na uporabnikovem računalniku, za samo informacijo o tem, ali je na računalnik nameščena neka določena pisava, pa je dovolj že podpora jeziku JavaScript.

Ti podatki so zelo uporabni za razlikovanje med računalniki, saj so seznami vtičnikov in pisav v velikem delu odvisni od potreb in preferenc posameznih uporabnikov.

Zakonitost sledenja z razločevanjem

Sledenje z razločevanjem je skladno z Zakonom o elektronskih komunikacijah - ta namreč omejuje le uporabo piškotkov in drugih tehnologij, ki shranjujejo podatke na uporabnikov računalnik. Informacijska pooblaščenka RS omenja "fingerprinting" v Smernicah o uporabi piškotkov [5] kot tehnologijo za prepoznavanje podpisa uporabnikove naprave oz. brskalnika. Pooblaščenka ne problematizira tega početja, opozarja pa na nujnost varovanja tudi na tak način pridobljenih osebnih podatkov.

2.2 *Profiliranje spletnih uporabnikov*

Profiliranje spletnih uporabnikov je zelo aktivno področje - tako z raziskovalnega, kot ekonomskega vidika. Velika spletna podjetja, kot so Google, Yahoo, Microsoft in ostali, razpolagajo z velikimi količinami podatkov o uporabnikih. Vedno večji delež njihovih prihodkov predstavlja spletno oglaševanje, zato vlagajo veliko truda v analizo obnašanja njihovih uporabnikov. Ta podjetja brez dvoma razpolagajo z najnaprednejšimi metodami za profiliranje, saj prinašajo že najmanjše izboljšave velike ekonomske koristi. Na žalost so metode za gradnjo profilov, še bolj pa podatki o uporabnikih, poslovna tajnost, zato lahko o podrobnostih njihovega delovanja le ugibamo.

Uporabniški profili se najpogosteje uporabljajo v sistemih za priporočanje [6, 7], pri personalizaciji spletnih aplikacij [8, 9], v učnih platformah [9] ipd. Kvaliteta profilov lahko močno vpliva na delovanje aplikacij, zato sta zbiranje podatkov o uporabnikih in gradnja uporabniških profilov zelo pomembna koraka pri gradnji takih aplikacij.

Struktura in vsebina uporabniškega profila je v prvi vrsti odvisna od podatkov, s katerimi razpolagamo. Konec prejšnjega stoletja, ko je oglaševanje postajalo vedno bolj pomemben dejavnik na spletu, so se za zbiranje informacij o uporabnikih pogosto uporabljale invazivne metode. Podjetja so vabila uporabnike k reševanju anket in vprašalnikov, z zbranimi informacijami pa so si potem pomagali pri segmentaciji populacije in ciljanem oglaševanju. Treba je povedati, da je bilo takrat spletno oglaševanje manj razširjeno in da so bili spletni uporabniki redkeje bombardirani z oglasi. Tu je verjetno treba iskati razlog, da so bili taki pristopi v tistem času relativno uspešni, saj je v današnjih časih popolnoma nerealno pričakovati, da bodo spletni uporabniki množično in z veseljem sodelovali v anketah ter oglaševalskim podjetjem brez razloga zaupali svoje osebne podatke, kot so na primer starost, spol in višina letnih prihodkov. Kljub temu, da dajejo spletni uporabniki vedno večji pomen zasebnosti, pa so se v zadnjem času za dober vir osebnih podatkov izkazale spletne nagradne igre.

Gradnja uporabniškega profila na podlagi vprašalnika je lahko zelo preprost način za izboljšavo uporabniške izkušnje. Petrelli [10] je s kratkimi vprašalniki zbiral informacije o starosti, spolu, poklicih, izobrazbi in navadah obiskovalcev muzeja. Izkazalo se je, da si z osebnimi podatki ni mogel veliko pomagati in da je za bolje izkoriščen obisk muzeja bolj pomembno vedeti, ali je obiskovalec prišel prvič, v kakšni skupini prihaja (družina, šola ali kot posameznik), kaj ga zanima in koliko časa ima na voljo. Slaba stran uporabe anket in vprašalnikov je dejstvo, da odgovorom uporabnikov ne moremo slepo verjeti. Njihovi odgovori že po definiciji ne morejo biti objektivni, še posebej pri bolj osebnih vprašanjih pa imamo lahko opravka tudi z neodkritostjo. V raziskavi [11] se je na primer pokazalo, da so študenti zelo precenjevali čas, ki ga preživijo na socialnem omrežju Facebook.

V zadnjem času se za profiliranje spletnih uporabnikov največkrat uporabljajo neinvazivne metode, ki veljajo za bolj prijazne do uporabnikov. Uporabnika in njegove aktivnosti na spletu ponavadi spremljamo dlje časa, potem pa zbrane informacije združimo v profil, ki modelira njegove interese. Takemu profilu pravimo tudi kontekstualni profil uporabnika. Pri gradnji profila iz uporabnikovega klikotoka (angl. clickstream) obravnavamo vsak obisk spletne strani kot indikator, da je vsebina spletne strani v skladu z njegovimi interesi. Vsebino vsake posamezne obiskane spletne strani navadno analiziramo s tehnikami podatkovnega rudarjenja. Thomas in sod. [12] predlagajo obogatitev spletnih strani z meta podatki po specifikacijah RDF[1], kar bi na elegan-

[1]Resource Description Framework http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

ten način izničilo potrebo po podatkovnem rudarjenju. Breme dodajanja oznak RDF spletnim vsebinam bi najverjetneje padlo na avtorje in urednike. Menimo, da se, vse dokler pomeni dodajanje oznak RDF za avtorje in urednike dodatno delo, to ne bo obneslo. Ta trend pa bi lahko spremenila dostopnost programskih knjižnic, ki bi to delo opravljale avtomatično, t.j. brez človeškega posredovanja.

Vivacqua [13] je razvila profilirno metodo i-ProSe, ki uporablja za gradnjo profila več virov informacij (življenjepisi, članki, vsebina elektronske pošte itd.). Profili imajo vektorsko obliko, se pravi, da držijo informacijo o pogostih ključnih besedah in pripadajočih utežeh. S profili brazilskih raziskovalcev je zgradila sistem za priporočanje, ki raziskovalcem predlaga možne projektne partnerje. V zadnjih letih so se zelo razširile prilagodljive platforme za e-izobraževanje [9]. Pri personalizaciji teh platform je potrebno vedeti predvsem, kakšne so želje, znanje in sposobnosti učencev. S spremljanjem aktivnosti posameznih učencev in njihovega napredka lahko platformo oz. naloge priredimo njihovim potrebam in tako izboljšamo učni proces.

Billsus in Pazzani [6] sta prepoznala pomen razlikovanja med uporabnikovimi dolgoročnimi in kratkoročnimi interesi. Pri gradnji sistema za priporočanje novic spletnim uporabnikom sta za vsakega uporabnika namesto enega zgradila dva modela uporabnikovih interesov. Z analizo zadnjih uporabnikovih interakcij s sistemom sta zgradila model uporabnikovih kratkoročnih interesov, za dolgoročne interese pa sta uporabila vse uporabnikove povratne informacije. S tem sta uspela sistem za priporočanje prilagoditi spreminjajočim se interesom spletnih uporabnikov in izboljšati kvaliteto priporočil.

Moderne metode za gradnjo profilov velikokrat uporabljajo dodatne vire informacij, kot so ontologije. Za večino problemov so dovolj dobre splošne ontologije, kot sta na primer WordNet[2] in DMOZ [14], za bolj specifične probleme pa obstajajo tudi domenske ontologije.

Eyharabide [15] je obogatila in izboljšala uporabniške profile s povezovalnimi pravili, znanja iz ontologije in razširjanja aktivacije (angl. spreading activation). Poskusi so bili izvedeni na vsega 52 študentih in zato nimajo velike vrednosti.

Middleton [16] je razvil dva sistema za priporočanje, s katerima je primerjal vektorske in ontološke uporabniške profile. Ugotovil je, da so seznami priporočil, zgrajeni z uporabo ontoloških profilov, kvalitetnejši, kar kaže na koristnost integracije ontologij v sisteme za priporočanje.

[2] WordNet http://wordnet.princeton.edu/

Beitzel [17] je v analizi podatkov spletnega iskalnika ugotovil, da je popularnost posameznih tematik zelo odvisna od ure v dnevu. Pri personalizaciji spletnih iskalnikov se lahko osredotočimo na uporabnikove kratkoročne [18] ali dolgoročne [8] interese. Tako Daoud [18] kot Sieg [8] sta si pri gradnji profilov uporabnikov pomagala z ontologijo "Open Directory Project" [14]. Daoud je s prepoznavanjem semantičnih preskokov razdelila klikotoke uporabnikov na seje, pri čemer se je za najboljše merilo izkazal Kendallov koeficient korelacije. Ob začetku vsake seje je ustvarila nov ontološki profil za uporabnika. Tekom seje je posodabljala uteži za posamezne koncepte v uporabnikovem profilu s t.i. razširjanjem ocen (angl. score propagation). Z razširjanjem ocen dodatno točkujemo koncepte, ki so v ontologiji povezani z neposredno točkovanimi. S poskusi na podatkovni množici HARD 2003 TREC [19] so pokazali, da njihov pristop statistično značilno izboljša natančnost iskanja. Sieg [8] je rezultate iskanja v spletnem iskalniku s ponovnim rangiranjem prilagodil uporabnikom na podlagi njihovih dolgoročnih interesov. Za točkovanje konceptov je uporabil razširjanje aktivacije (angl. spreading activation), ki deluje na podoben način kot prej omenjeno razširjanje ocen.

Algoritem WebDCC [20] uporablja razvrščanje konceptov za izdelavo modela uporabnikovih interesov. Godoy je razvila metodo za gradnjo ontoloških profilov uporabnikov [21], pri kateri je algoritmu WebDCC dodala še časovno pozabljanje, na podlagi pozitivnih povratnih informacij pa je še dodatno zvišala ocene ustreznim konceptom. V poskusih so simulirali uporabnike s spreminjajočimi se interesi, rezultati pa so pokazali, da se lahko tako zgrajeni profili hitro prilagajajo spremembam v obnašanju uporabnikov.

3

Nova profilirna metoda AverageActionFC

3.1 Metoda AverageActionFC

Definicija 3: Ontologija opisuje množico konceptov v neki domeni in odnose med njimi. Največkrat je predstavljena kot graf, v katerem so koncepti predstavljeni z vozlišči, odnosi pa s povezavami med njimi.

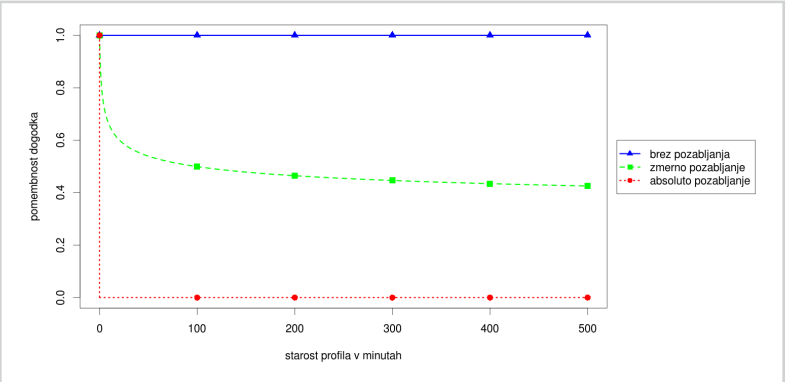
Definicija 4: Ontološki profil uporabnika je model uporabnikovih interesov, pri katerem je za predlogo uporabljena splošna ali domenska ontologija. Profil je sestavljen iz ontoloških konceptov, vsakemu pa je pripisana tudi utež, s katero označujemo pomembnosti posameznih konceptov.

Definicija 5: Klikotok (angl. clickstream) je zaporedje uporabnikovih spletnih aktivnosti (obiski spletnih strani), ki se navadno beležijo v dnevnike spletnih strežnikov. Vsak vnos je navadno opremljen s podatki, kot so čas dogodka, naslov URL obiskane spletne strani ipd.

Obstoječe metode za gradnjo ontoloških profilov (glej razdelek 2.2) temeljijo na analizi preteklih uporabnikovih aktivnosti. Običajno se za predlogo uporabi ena od splošnih ontologij, npr. DMOZ (the Open Directory Project) [14]. Na obiskanih spletnih straneh poskušajo detektirati koncepte iz ontologije, nakar jim povišajo oceno pomembnosti. Nekatere metode prilagajajo profile spreminjajočim se interesom uporabnikov:

- z detekcijo semantičnih preskokov v zaporedju uporabnikovih aktivnosti (metoda avtorice Daoud [18]) ali
- s tehniko časovnega pozabljanja (npr. metoda avtorice Godoy [21]), s katero dajo pri gradnji profila večji pomen konceptom, ki so bili detektirani nazadnje.

Glavna pomanjkljivost obstoječih metod je njihova omejenost na podatke iz uporabnikovega klikotoka, ki je velikokrat kratek ali enoličen ter predstavlja zelo omejen vir informacij. To omejitev skušajo nekatere metode preseči z uporabo razširjanja ocen (metoda avtorice Daoud [18]) ali razširjanja aktivacije (metoda avtorja Sieg [8]), ki na



Slika 3.1

Modra krivulja predstavlja pomembnost posameznih dogodkov brez pozabljanja - vsi dogodki so enako pomembni. Z rdečo krivuljo je prikazano absolutno pozabljanje, pri katerem pozabimo na vse dogodke razen zadnjega. Zelena krivulja predstavlja zmerno pozabljanje, pri katerem pomembnost dogodka pada s časom, ki je pretekel. Uporabljena je funkcija logaritma (3.1) s hitrostjo pozabljanja $a = 100$.

podlagi strukture uporabljene ontologije razširjata ocene konceptov tudi na koncepte, ki niso bili detektirani v obiskanih spletnih straneh. Nova metoda AverageActionFC uporablja za gradnjo profila za razliko od obstoječih metod poleg uporabnikovega klikotoka še informacije v profilih njemu podobnih uporabnikov.

V nadaljevanju sta predstavljeni dve tehniki, ki sta del nove metode za gradnjo ontoloških profilov: časovno pozabljanje in popravljanje profilov s prototipnimi profili.

3.1.1 Časovno pozabljanje

Tehnika časovnega pozabljanja se pogosto uporablja tako na področju profiliranja uporabnikov kot tudi v družboslovnih znanostih. Osnova te tehnike je padajoča funkcija, s katero modeliramo pozabljanje ali počasno spreminjanje interesov. V literaturi so omenjene linearne, polinomske, eksponentne, logaritmične in sigmoidne funkcije. V začetnih poskusih se je izkazalo, da je izmed naštetih za modeliranje pozabljanja na podatkih oglaševalske mreže najbolj primerna funkcija logaritma.

V enačbi (3.1) označujemo pomembnost uporabnikove akcije z $importance_{action}$, definirana pa je na podlagi časa age_{action}, ki je pretekel od takrat in hitrosti pozabljanja, ki jo določa osnova logaritma a. Na sliki 3.1 so prikazane tri hitrosti pozabljanja in njihov vpliv na izračun pomembnosti uporabnikovih aktivnosti.

$$importance_{action} = \frac{1}{1 + log_a(age_{action} + 1)} \tag{3.1}$$

3.1.2 Popravljanje profilov s prototipnimi profili

Glavna prednost nove metode za profiliranje uporabnikov je izboljšava uporabnikovega profila z vnosom domenskega znanja, ki je predstavljeno kot množica prototipnih profilov. Vsak od prototipnih profilov predstavlja model interesov tipičnega uporabnika oz. skupine uporabnikov, vsi prototipni profili skupaj pa pokrivajo celotno ali pa večino populacije uporabnikov.

Popravljanje uporabnikovega profila (glej Algoritem 2) poteka tako, da v prvem koraku v množici prototipnih profilov najdemo tistega, ki je uporabnikovemu najbolj podoben. V drugem koraku približamo ocene konceptov v uporabnikovem profilu ocenam istih konceptov v prototipnem profilu. Mero popravljanja imenujemo utež za popravljanje $w_{correct}$, zavzema pa lahko poljubne vrednosti med $0 \leq w_{correct} \leq 1$. V primeru, da uporabimo vrednost uteži za popravljanje $w_{correct} = 0$, je popravljeni profil enak prvotnemu, z utežjo $w_{correct} = 1$ pa izenačimo uporabnikov profil s prototipnim.

Prototipni uporabniški profili

Množico prototipnih profilov lahko določi domenski strokovnjak, ki pozna interese in obnašanje ciljne populacije, lahko pa jih definiramo na podlagi analize preteklih aktivnosti celotne populacije uporabnikov. V kolikor imamo opravka z novo aplikacijo, so lahko podatki o uporabnikih preskopi, da bi lahko z analizo prišli do stabilne in zanesljive množice prototipov. Če želimo uporabljati popravljanje profilov s prototipi, je tako potrebna pomoč strokovnjakov. Če pa imamo na razpolago večjo količino podatkov o preteklih aktivnostih uporabnikov, lahko z analizo teh podatkov definiramo množico prototipov, ki bo koristen vir informacij za popravljanje uporabniških profilov.

V poskusih, ki so predstavljeni v naslednjem poglavju, smo imeli na voljo klikotoke večjega števila spletnih uporabnikov, ki se raztezajo skozi daljše časovno obdobje. To obdobje smo razdelili na dva dela:

1. podatki za kalibracijo in določitev prototipnih profilov ter
2. podatki za evalvacijo metod za profiliranje uporabnikov.

Klikotoke uporabnikov iz prvega obdobja smo uporabili za gradnjo preprostih ontoloških profilov. Z algoritmom AverageAction (Algoritem 1) smo zgradili uporabnikov

profil tako, da smo za vsakega od konceptov iz ontologije izračunali povprečje njegovih ocen v vseh dogodkih iz uporabnikovega klikotoka. Na profilih, izračunanih z algoritmom AverageAction, smo izvedli razvrščanje. Uporabili smo standardni algoritem k-means [22], katerega rezultat je množica centroidov. Vsak od tako izračunanih centroidov imenujemo prototipni profil, saj predstavlja interese večjega števila uporabnikov.

Algoritem 1: Algoritem za gradnjo ontoloških uporabniških profilov AverageAction.

```
Input: ontology, actions
Output: profile
// inicializacija profila
foreach category do
    profile[category] ← 0
// gradnja profila iz uporabnikovih preteklih aktivnosti
foreach action do
    foreach category do
        profile[category] ← profile[category] + action[category]
// normalizacija ocen konceptov v profilu
foreach category do
    profile[category] ← profile[category]/actions.length
```

V podatkovnem rudarjenju se za mero podobnosti med dvema vektorjema pogosto uporablja kosinusna podobnost [23], ki je definirana kot kosinus kota med njima. Uporaba kosinusne podobnosti je smiselna, če velja predpostavka, da sta vektorja med seboj neodvisna. Pri računanju podobnosti med ontološkimi profili ta predpostavka ne velja - koncepti v ontologiji so namreč med seboj povezani ravno na podlagi medsebojne sorodnosti. To omejitev presega *posplošena kosinusna podobnost*, kot jo je definiral Ganesan [24]. Pri posplošeni kosinusni podobnosti je produkt med vhodnima vektorjema A in B (enačba 3.3) definiran kot vsota vseh produktov komponent obeh vektorjev $a_i \cdot b_j$, pri čemer se sorodnost med obema vektorjema upošteva preko produkta členov $\vec{l_i} \cdot \vec{l_j}$ (enačba 3.4), ki je definiran s pomočjo funkcij:

- $LCA(l_i, l_j)$ (angl. lowest common ancestor) - najnižji skupni prednik vozlišč l_i in l_j ter
- $depth(l_i)$ - globina vozlišča l_i v ontologiji.

V izračunu podobnosti med dvema profiloma (glej enačbe 3.2-3.4) so vključeni tako točkovanja konceptov v obeh profilih kot tudi struktura ontologije, zato smo pri razvrščanju ontoloških uporabniških profilov z algoritmom k-means za mero podobnosti uporabili posplošeno kosinusno podobnost (GCSM).

$$sim_{GCSM}(A,B) = \frac{\vec{A}\cdot\vec{B}}{\sqrt{\vec{A}\cdot\vec{A}}\cdot\sqrt{\vec{B}\cdot\vec{B}}} \tag{3.2}$$

$$\vec{A}\cdot\vec{B} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \cdot b_j \cdot \vec{l_i}\cdot\vec{l_j} \tag{3.3}$$

$$\vec{l_i}\cdot\vec{l_j} = \frac{2\cdot depth(LCA(l_i,l_j))}{depth(l_i)+depth(l_j)} \tag{3.4}$$

Zelo znan problem pri razvrščanju z algoritmom k-means je iskanje optimalnega števila gruč k. Optimum je odvisen predvsem od podatkov, ki jih uporabljamo za gradnjo profilov, zato smo ga določili s kalibracijo.

3.1.3 *Profilirni algoritem AverageActionFC*

Pri razvoju algoritma AverageActionFC (končnica imena algoritma izhaja iz angleških besed *F*orgetting in *C*orrection), ki je izboljšava algoritma AverageAction (glej Algoritem 1), smo bili osredotočeni predvsem na izboljšanje robustnosti in fleksibilnosti algoritma.

Mehanizem časovnega pozabljanja omogoča prilagajanje profila spreminjajočim se interesom uporabnika, s prilagajanjem hitrosti pozabljanja pa lahko optimiziramo uporabnikov profil tako, da bolje modelira njegove kratkoročne ali dolgoročne interese.

Robustnost profila v smislu, da gradi relativno kvalitetne uporabniške profile tudi, ko je na voljo zelo malo podatkov o uporabnikovih preteklih aktivnostih, smo zagotovili s tehniko popravljanja profila s prototipi. V praksi je pomanjkanje podatkov pogost pojav, zato je pomembno, da se pri gradnji profila uporabijo tudi drugi viri podatkov. S približevanjem uporabnikovega profila prototipu vnašamo v profil podatke o interesih njemu podobnih uporabnikov, kar lahko občutno zviša kvaliteto tako zgrajenega profila.

Algoritem AverageActionFC (glej Algoritem 2) ima naslednje argumente:

- *ontology* - referenčna ontologija, ki služi kot predloga za uporabniške profile,
- *actions* - seznam uporabnikovih preteklih aktivnosti,
- f_{forget} - funkcija časovnega pozabljanja,
- *prototypes* - seznam prototipnih profilov in
- $w_{correct}$ - utež za popravljanje profila.

Dodatna predpostavka pri podatkih o uporabnikovih preteklih aktivnostih je, da se da vsako posebej semantično preslikati v referenčno ontologijo.

V prvem koraku gradnje profila uporabimo funkcijo časovnega pozabljanja za določanje pomembnosti posameznih aktivnosti iz uporabnikovega klikotoka, nakar zapišemo ocene h konceptom ontologije. Drugi korak algoritma je popravljanje profila s prototipi. V množici prototipnih profilov moramo najprej najti tistega, ki je najbolj podoben uporabnikovemu profilu, potem pa približamo ocene konceptov v uporabnikovem profilu ocenam v prototipnem v skladu z vrednostjo uteži za popravljanje $w_{correct}$.

V nadaljevanju so za nekatere različice algoritma AverageActionFC uporabljena druga imena:

- *AverageAction* - Najpreprostejši algoritem za gradnjo profilov (glej Algoritem 1), ki pri svojem delovanju ne uporablja niti časovnega pozabljanja niti popravljanja s prototipi. Ocene konceptov v končnem profilu so enake povprečnim ocenam teh profilov v dogodkih iz učnega klikotoka.
- *AverageActionF* - Ta različica uporablja mehanizem časovnega pozabljanja, ne uporablja pa popravljanja profilov s prototipi. Hitrost pozabljanja je zmerna, osnova logaritma a v enačbi 3.1 običajno zavzema vrednosti $10 \leq a \leq 100$.
- *LastAction* - Pri tej različici je uporabljena najvišja hitrost pozabljanja, kar pomeni, da algoritem zanemari vse dogodke v učnem klikotoku razen zadnjega. Uporabnikov profil vsebuje enake ocene konceptov kot zadnji dogodek iz učnega klikotoka.

Algoritem 2: Algoritem za gradnjo ontoloških profilov AverageActionFC, ki uporablja tehniki časovnega pozabljanja in popravljanja profilov s prototipi.

Input: *ontology*, *actions*, f_{forget}, *prototypes*, $w_{correct}$
Output: *profile*
`// inicializacija profila`
foreach category do
  $profile[category] \leftarrow 0$
`// gradnja osnovnega profila iz uporabnikovih preteklih aktivnosti z uporabo funkcije časovnega pozabljanja`
$w_{sum} \leftarrow 0$
foreach action do
  $w_{forget} \leftarrow f_{forget}(age_{action})$
  foreach category do
    $profile[category] \leftarrow profile[category] + w_{forget} * action[category]$
  $w_{sum} \leftarrow w_{sum} + w_{forget}$
`// normalizacija ocen konceptov v profilu`
foreach category do
  $profile[category] \leftarrow profile[category]/w_{sum}$
`// gradnja končnega profila z uporabo popravljanja s prototipi`
$prototype \leftarrow find\_most\_similar(prototypes, profile)$
foreach category do
  $x_{typ} \leftarrow w_{correct} * prototype[category]$
  $profile[category] \leftarrow (1 - w_{correct}) * profile[category] + x_{typ}$

3.2 Ostale razvite profilirne metode

V zadnjih letih smo razvili še nekaj povsem drugačnih metod za gradnjo ontoloških profilov. Pri gradnji se nismo omejili le na gradnjo statičnih profilov, ki vsebujejo le tako ali drugače dobljene ocene posameznih konceptov iz ontologije. Edino pravo vrednost uporabnikovega profila smo videli namreč v njegovi zmožnosti napovedovanja uporabnikovih interesov, zato smo dostikrat v uporabnikov profil namesto statičnih ocen konceptov vgradili katerega od preprostejših napovednih modelov, zgrajenega z algoritmi strojnega učenja.

3.2.1 Popravljanje profilov s časovnimi statistikami

S popravljanjem uporabniških profilov s časovnimi statistikami prilagajamo ocene konceptov v uporabnikovem profilu glede na trenutno popularne tematike. Tako lahko v sicer statičen profil, ki je zgrajen na podlagi uporabnikovih preteklih aktivnosti, vnesemo določeno mero dinamičnosti.

Časovne statistike

Kot dodaten vir informacij za izboljšanje uporabniških profilov smo uporabili podatke o popularnosti posameznih tematik v preteklosti, ki smo jih imenovali *časovne statistike*. Pri izračunu teh statistik smo uporabili vse obiske spletnih strani iz klikotokov, ki so bili sicer uporabljani za kalibracijo metod. Izračunali smo navadno povprečje uteži konceptov po vseh dogodkih v uporabniških klikotokih, tako da statistike dejansko odražajo popularnost posameznih tematik ob določenih časovnih obdobjih.

Pri izračunu smo uporabili 4 ravni podrobnosti:

1. dnevne statistike za posamezne dneve v tednu: 7 časovnih statistik,
2. dnevne statistike za delavnike in vikende: 2 časovnih statistiki,
3. urne statistike za posamezne dneve v tednu: $7 * 48 = 168$ časovnih statistik in
4. urne statistike za delavnike in vikende: $2 * 24 = 48$ časovnih statistik

Pregled omenjenih statistik je pokazal, da se pri nekaterih tematikah njihova popularnost ne spreminja dosti s časom – take so na primer izobraževanje, telekomunikacije, tabloidi, zabava in video igre. Pri drugih tematikah pa smo opazili, da se njihova popularnost očitno poveča v določenih časovnih obdobjih:

- med tednom ponoči (2h-5h): računalniška strojna oprema, mobilni telefoni, video naprave in ostala potrošniška elektronika,
- med tednom zjutraj in v dopoldanskih urah (6h-13h): poslovanje, upravljanje s človeškimi viri, davki, transport in logistika,
- med tednom v času malic in kosil (10h-15h): hrana, kuhanje in recepti,
- med tednom popoldne (16h-20h): igrače, stripi in risanke,
- med vikendi ponoči (3h-6h): šport,
- med vikendi čez cel dan (9h-20h): igrače, kuhanje, recepti,
- med vikendi dopoldne (9h-13h): stripi in risanke,
- med vikendi popoldne (12h-18h): filmi,
- med vikendi popoldne in zvečer (15h-22h): domači ljubljenčki in
- vsak dan ponoči (3h-7h): igre na srečo.

Popravljanje profilov s časovnimi statistikami

Postopek popravljanja uporabnikovega profila s časovnimi statistikami je podoben popravljanju s prototipnimi profili v algoritmu AverageActionFC (glej Algoritem 2 na strani 26). Za razliko od algoritma AverageActionFC, pri katerem se popravljanje profila zgodi takoj po gradnji osnovnega uporabnikovega profila, se zgodi popravljanje profila s časovnimi statistikami vsakič, ko je potrebno izdelati napoved uporabnikovih interesov. Pri vsaki napovedi se na podlagi trenutnega časa izbere ustrezna časovna statistika, popravljanje ocen konceptov v uporabnikovem profilu pa se v skladu s kalibrirano vrednostjo $w_{correct}$ in izbrano statistiko izvede na enak način kot pri AverageActionFC.

Ker je potrebno osnovni uporabnikov profil popravljati za vsako napoved, je koristno pred popravljanjem narediti kopijo osnovnega profila, ki se jo lahko potem uporablja za nadaljnje napovedi.

4

Poskusi: Gradnja in evalvacija uporabniških profilov

4.1 Metodologija testiranja metod za gradnjo ontoloških profilov

Razvite metode za gradnjo ontoloških profilov smo primerjali z obstoječimi metodami avtorjev Daoud [18], Godoy [21] in Sieg [8] (glej razdelka 2.2 in 3.1). Nekatere od primerjanih metod je potrebno pred uporabo kalibrirati, zato smo podatkovno množico najprej razdelili na dva dela:

1. podatki za *kalibracijo* metod - ti podatki so služili izključno iskanju optimalnih vrednosti argumentov in

2. podatki za *medsebojno primerjavo* metod - na podlagi teh podatkov smo izvajali evalvacijo in primerjavo metod, pri čemer smo v posameznih metodah vedno uporabili s kalibracijo dobljene vrednosti parametrov.

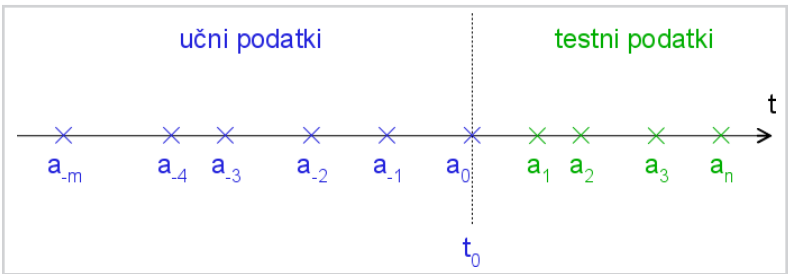
Z razdelitvijo podatkov na kalibracijske in primerjalne smo se izognili problemu prevelikega prileganja učnim podatkom (angl. overfitting) in povečali zaupanje v nepristranost dobljenih rezultatov. V veliko primerih so v obeh množicah podatkov zastopani isti uporabniki, vendar pa nobena od metod te informacije pri svojem delovanju ne izkorišča.

Pri zasnovi metodologije za evalvacijo in primerjavo metod za profiliranje spletnih uporabnikov so nas vodila naslednja vprašanja:

- Kako oceniti kvaliteto uporabnikovega profila?
- Koliko časa je zgrajeni profil še aktualen oz. koliko časa dobro modelira uporabnikove interese?
- Kakšen vpliv ima na kvaliteto profila količina informacij, ki jo uporabimo pri gradnji profila?
- Ali je metoda za profiliranje A boljša od metode B?

4.1.1 Kvaliteta uporabnikovega profila

Pri ocenjevanju kvalitete uporabnikovega profila smo se osredotočili na to, kako dobro modelira uporabnikove interese, se pravi kakšna je vrednost profila kot napovednega modela. Uporabnikov klikotok smo zato razdelili na učne in testne dogodke (slika 4.1). Učne dogodke smo uporabili za gradnjo uporabnikovega profila, ki ga potem



Slika 4.1

Prikaz razdelitve uporabnikovega klikotoka, ki vsebuje $m + n + 1$ dogodkov, na učne in testne podatke. Prvih $m + 1$ dogodkov v klikotoku se uporabi za gradnjo uporabnikovega profila, zadnjih n dogodkov pa za ocenjevanje kvalitete profila. Dogodek a_0 je najnovejši dogodek v delu klikotoka, ki se uporabi za gradnjo profila, čas tega dogodka t_0 pa obravnavamo kot čas gradnje profila. Času, ki preteče od gradnje profila do testnega dogodka, pravimo starost profila.

nismo več posodabljali, kvaliteto uporabnikovega profila pa smo ocenjevali na podlagi primerjave med profilom in testnimi dogodki v uporabnikovem klikotoku.

Za oceno kvalitete uporabnikovega profila Q_u smo uporabili posplošeno kosinusno podobnost (GCSM) med profilom in testnim dogodkom. Ker smo lahko kvaliteto profila ocenjevali le ob časih testnih dogodkov, smo na podlagi razlike med časom gradnje profila in časom testnega dogodka definirali kvaliteto uporabnikovega profila ob starosti profila t kot $Q_u[t]$.

$$Q_u[t] = sim_{GCSM}(profile, content_{visit}) \tag{4.1}$$

Z združitvijo ocen kvalitet profilov večjega števila uporabnikov smo definirali povprečno kvaliteto profilov v določenem obdobju $Q_{avg}[t], t \in [t_{min}, t_{max}]$ (enačba 4.2). S to oceno lahko učinkovito merimo uspešnost metode za profiliranje in jo primerjamo z ostalimi metodami.

$$Q_{avg}[t] = \frac{\sum_{u \in users} Q_u[t]}{n_{events}[t]}, t \in [t_{min}, t_{max}] \tag{4.2}$$

4.1.2 Vpliv starosti profila in količine informacij na kvaliteto profila

Običajno velja predpostavka, da ima starost negativen vpliv na kvaliteto uporabnikovega profila. S spremljanjem uporabnika in sprotnim posodabljanjem njegovega profila bi v teoriji lahko ta vpliv izničili, vendar pa je v praksi, še posebej ob urah, ko je na

spletu največ uporabnikov, to lahko prevelik zalogaj za razpoložljivo strojno opremo. V takih primerih se je potrebno vprašati, če je res smiselno posodabljati uporabnikov profil z vsakim novim dogodkom, ki se pojavi v njegovem klikotoku. Pogostost posodabljanja uporabnikovega profila je tako vprašanje razpoložljivosti strojne opreme, hitrosti metode za profiliranje in hitrosti padanja kvalitete profila s časom, lahko pa je to tudi povsem poslovna odločitev. V vsakem primeru je koristno vedeti, kako dobro lahko po določenem času še modeliramo uporabnikove interese, zato smo povprečne ocene kvalitet profilov razdelili na 11 časovnih obdobij, ki se raztezajo od trenutka gradnje profila do starosti profila 4 tedne:

- $0\ s \leq starost < 3\ s$,
- $3\ s \leq starost < 1\ min$,
- $1\ min \leq starost < 10\ min$,
- $10\ min \leq starost < 30\ min$,
- $30\ min \leq starost < 1\ ura$,
- $1\ ura \leq starost < 2\ uri$,
- $2\ uri \leq starost < 6\ ur$,
- $6\ ur \leq starost < 12\ ur$,
- $12\ ur \leq starost < 1\ dan$,
- $1\ dan \leq starost < 1\ teden$,
- $1\ teden \leq starost < 4\ tedni$.

Gibanje povprečne kvalitete profila smo spremljali tudi glede na števila učnih dogodkov, ki so bili uporabljeni za gradnjo profila. Predvidevali smo, da lahko pri kratkih učnih klikotokih vsaka dodatna informacija pomaga izboljšati kvaliteto profila, po gradnji profila z uporabo velikega števila učnih dogodkov pa je profil težko izpopolniti. Izvajali smo več serij poskusov, v vsaki seriji pa smo nastavili fiksno dolžino učnih klikotokov. Uporabljali smo učne klikotoke dolžin od 5 do 100 učnih dogodkov, s primerjavo rezultatov pa smo lahko ugotovili, kako število dogodkov, ki so bili uporabljeni za gradnjo profila, vpliva na kvaliteto profilov.

4.1.3 Medsebojna primerjava metod za profiliranje

Da bi bila medsebojna primerjava metod čim bolj objektivna, so bili pri evalvaciji vsake metode uporabljeni isti podatki. To pomeni, da so bili uporabniški profili zgrajeni z istimi klikotoki, prav tako pa so bili za ocenitev kvalitete zgrajenih profilov vedno uporabljeni isti testni dogodki. Da bi lahko čimbolj zanesljivo odgovorili na vprašanje, ali je ena metoda za profiliranje boljša od druge, smo uporabili statistične teste.

Za vsako od primerjanih metod smo zgradili seznam ocen kvalitet profilov, pri čemer so vsi primerjani seznami enako dolgi, za enakoležne ocene pa so bili uporabljeni isti testni dogodki. Za primerjavo smo uporabili parni Wilcoxonov test predznačenih rangov [25], ki za razliko od Studentovega t-testa ne predpostavlja normalne porazdelitve podatkov.

4.2 Poskusi: Kvaliteta profilov uporabnikov spletne oglaševalske mreže Httpool

V spletno oglaševalsko mrežo je vključeno večje število založnikov in oglaševalcev, oglaševalska mreža pa deluje kot posrednik.

Oglaševalci posredujejo mreži svoje oglase, mreža pa jim omogoča lažji dostop do spletne populacije. S ciljnim oglaševanjem lahko oglaševalska mreža prikazuje oglase le spletnim uporabnikom, za katere se predvideva, da jih taka vsebina zanima. Spletni založniki dajejo dele svoji spletnih strani v najem oglaševalski mreži, v zameno za oglasni prostor pa dobijo od mreže denarno nadomestilo. V kolikor je v oglaševalsko mrežo vključenih dovolj različnih oglaševalcev in spletnih založnikov, lahko mreža s pametnim prikazovanjem oglasov občutno izboljša uspešnost oglaševanja, kar se odraža na višjih vrednosti CTR in bolj ekonomični porabi oglaševalskega denarja.

Definicija 6: *CTR* ali "click-through rate" je delež prikazanih oglasov, na katere so uporabniki kliknili (glej enačbo 4.3). To je najpogosteje uporabljena mera za uspešnost posameznih oglasov in oglaševalskih kampanj, ne odraža pa vedno ekonomske uspešnosti, saj ne upošteva različnih cen klikov na oglase in stroškov oglaševanja.

$$CTR = \frac{num_{clicked}}{num_{displayed}} \tag{4.3}$$

Za potrebe pametnega oglaševanja zbira oglaševalska mreža podatke o:

- vsebinah spletnih strani,
- vsebinah oglasov in
- aktivnostih spletnih uporabnikov.

4.2.1 Opis podatkov oglaševalske mreže

Glavna vira podatkov so strežniški dnevniki oglaševalske mreže in podatkovna baza kategoriziranih vsebin spletnih strani.

Strežniški dnevniki

Strežniški dnevniki vsebujejo podatke o ogledih spletnih strani, ki so vključene v mrežo, prikazanih oglasih in klikih na oglase, ki so se zgodili v okviru oglaševalske mreže. Za gradnjo in evalvacijo ontoloških profilov spletnih uporabnikov smo uporabili le oglede spletnih strani. Poleg časa obiska in naslova URL obiskane spletne strani vsebujejo dnevniki tudi številko ID spletnega uporabnika in nekatere podatke zahtevka HTTP.

V poskusih so bili uporabljeni podatki iz obdobja od avgusta do novembra 2011. V tistem obdobju še ni veljal novi Zakon o elektronskih komunikacijah [3], zato smo spletnim uporabnikom lahko sledili s pomočjo spletnih piškotkov.

Vsebino strežniških dnevnikov smo razdelili na:

- kalibracijske podatke (avgust in september 2011) in
- podatke za primerjavo profilirnih metod (oktober in november 2011).

Iz strežniških dnevnikov smo izluščili klikotoke posameznih spletnih uporabnikov, pri čemer smo se omejili le na vnose, ki opisujejo obiske spletnih strani. Vsak vnos v strežniškem dnevniku opisuje en uporabnikov obisk spletne strani. Za gradnjo in evalvacijo uporabniških profilov smo uporabili naslednje podatke:

- čas obiska spletne strani,
- naslov URL obiskane strani in
- vsebino vrstice “User-Agent”, ki je del glave zahtevka HTTP.

Vrstica "User-Agent" vsebuje podatke o napravi, ki je poslala zahtevek HTTP na strežnik. Z uporabo knjižnice WURFL [26] lahko iz vsebine te vrstice ugotovimo, s katerim operacijskim sistemom in brskalnikom je uporabnik obiskal spletno stran. To nam je omogočilo, da smo lahko iz množice klikotokov izločili spletne robote, preostale uporabnike pa smo razdelili na:

- *namizne spletne uporabnike*, ki uporabljajo namizne in prenosne računalnike (operacijski sistemi Microsoft Windows, Linux, OS X, itd.) ter
- *mobilne uporabnike*, ki uporabljajo tablične računalnike in mobilne telefone (operacijski sistemi Android, iOS, Windows Mobile, itd.).

V tabeli 4.1 je prikazano število evalvacij uporabniških profilov, na podlagi katerih so pridobljeni rezultati, prikazani v tem poglavju.

Tabela 4.1

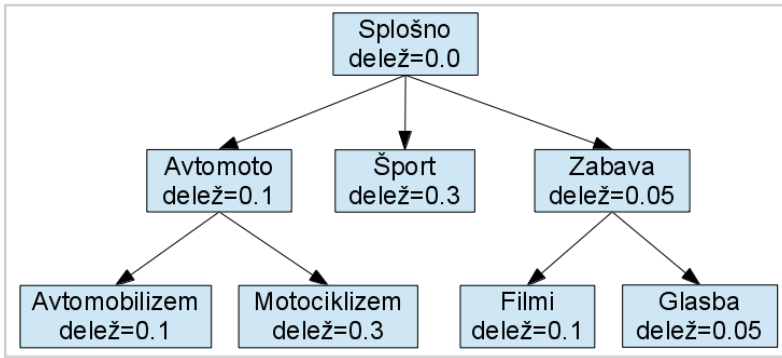
Število evalvacij uporabniških profilov v poskusih na podatkih oglaševalske mreže.

populacija	kalibracija	primerjava metod
namizni uporabniki	1.966.684	1.633.132
mobilni uporabniki	597.789	766.521

Kategorizacija vsebin spletnih strani

Oglaševalska mreža vzdržuje bazo naslovov URL, na katerih se prikazujejo njeni oglasi. Spletni roboti, ki so del oglaševalske mreže, obiskujejo te spletne strani in analizirajo njihovo vsebino. Proces kategorizacije spletnih vsebin, ki je opisan v tem poglavju, sem kot programski arhitekt v podjetju Httpool razvil leta 2008 in se od takrat naprej uporablja z potrebe kontekstualnega oglaševanja na evropskih in azijskih trgih. Trenutno je podprtih 15 evropskih jezikov, med katerimi je seveda tudi slovenščina.

Oglaševalska ontologija. Pri analizi vsebin spletnih strani se uporablja v podjetju razvita ontologija, ki vsebuje 130 tematskih kategorij, organiziranih v drevesno obliko (manjši izsek ontologije je prikazan na sliki 4.2).



Slika 4.2

Prikazan je manjši izsek oglaševalske ontologije, ki se uporablja kot orodje za kategorizacijo vsebin spletnih strani in kot predloga za profile spletnih uporabnikov v oglaševalski mreži. V korenu drevesa je kategorija "Splošno", ostala vozlišča pa predstavljajo bolj konkretne tematske kategorije. Vsako vozlišče v drevesu nosi tudi utež za pripadajočo kategorijo, ki ponazarja njen delež v vsebini spletne strani oz. njeno pomembnost v uporabnikovem profilu.

Vsako vozlišče v drevesu predstavlja eno tematsko kategorijo:

- *koren drevesa* predstavlja kategorijo "Splošno",
- *ostala vozlišča* pa predstavljajo bolj specifične tematike - čim bolj kot je vozlišče oddaljeno od korena drevesa, tem bolj specifično kategorijo predstavlja.

Proces kategorizacije spletnih vsebin deluje v naslednjih korakih:

1. *Ekstrakcija besedila* iz spletnih dokumentov,
2. Predobdelava besedila z *lematizacijo*. Na podlagi slovničnih pravil se vsem besedam odstranijo nepomembne predpone in pripone.
3. *Iskanje ključnih besed in fraz* v besedilu. Za vsako od tematskih kategorij v oglaševalski ontologiji je določen seznam od 30 do 100 ključnih besed in fraz, za katere se pričakuje, da se pojavljajo v tovrstnih besedilih. Nekatere ključne besede se lahko zaradi sorodnosti med kategorijami in večpomenskosti besed pojavljajo v različnih kategorijah.
4. *Prisotnost posameznih tematik v besedilu* $score_{category}$ se določa glede na kosinusno podobnost med besedilom in seznamom ključnih besed za vsako tematsko kategorijo, pri čemer je besedilo predstavljeno kot vreča besed (angl. bag-of-words). Izračun kosinusne podobnosti smo prilagodili tako, da so večbesedne

fraze zaradi večje specifičnosti in redkosti dodatno točkovane. Fraza "prva pomoč" namreč zelo očitno namiguje na zdravstveno tematiko, posamezni besedi "prva" in "pomoč" pa sami po sebi ne nosita velike sporočilne vrednosti.

5. Z *normalizacijo rezultatov* poskrbimo, da je vsota deležev vseh tematskih kategorij enaka $\sum score_{category} = 1.0$.

4.2.2 Kalibracija metode AverageActionFC

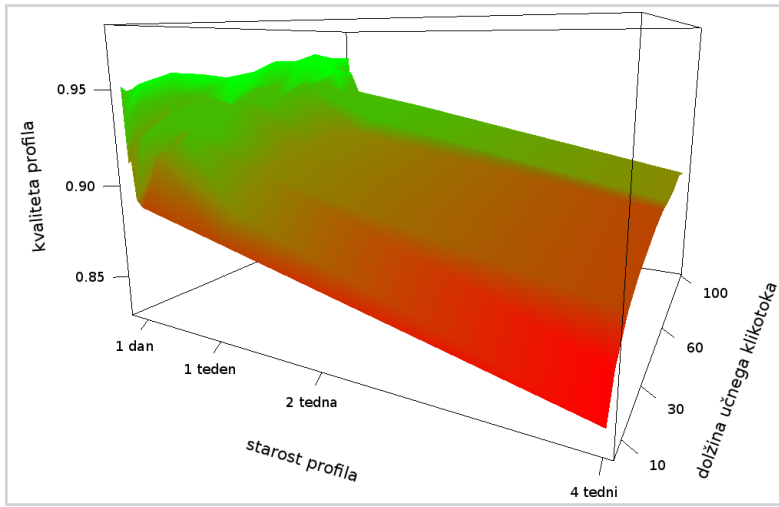
Metoda AverageActionFC (Algoritem 2) zahteva kalibracijo treh parametrov:

1. hitrosti pozabljanja,
2. množice prototipnih profilov in
3. uteži $w_{correct}$ za popravljanje profila.

Hitrost pozabljanja v metodi AverageActionFC lahko prilagajamo preko osnove logaritma v enačbi 3.1, ki je označena kot a. Uporaba manjših vrednosti parametra a se prevede v hitrejše pozabljanje preteklih dogodkov, kar pomeni, da dajemo večji pomen bolj nedavnim dogodkom. Pri visokih vrednostih parametra a je časovno pozabljanje manj intenzivno, pri teoretični vrednosti $a = \infty$ pa so si vsi dogodki iz uporabnikovega klikotoka enakovredni. Sliki 4.3 in 4.4 prikazujeta gibanje kvalitete uporabniških profilov namiznih in mobilnih uporabnikov glede na starost profilov in dolžino učnih klikotokov. Za gradnjo profilov je bil uporabljen algoritem AverageAction, ki ne uporablja časovnega pozabljanja.

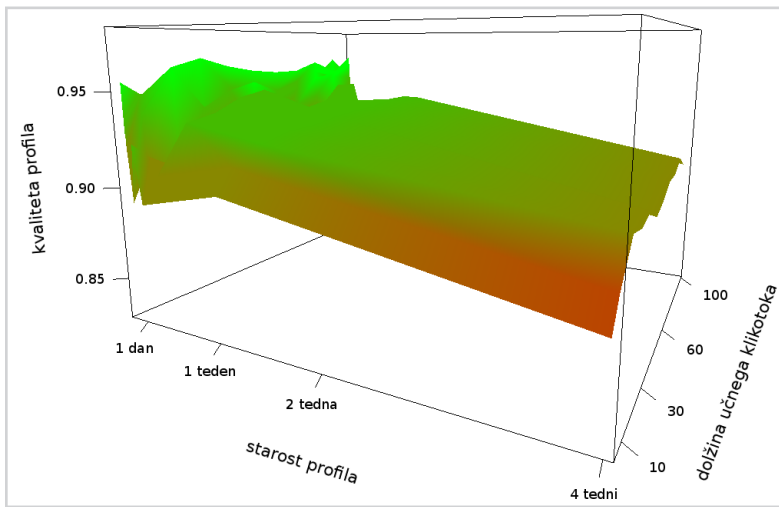
Pri kalibraciji hitrosti pozabljanja smo za kriterij primernosti parametra a vzeli kvaliteto napovedi uporabnikovih srednjeročnih interesov, pri čemer smo to obdobje definirali s starostjo profila $1\ ura \leq starost < 2\ uri$. Za gradnjo profilov smo uporabili dolge ($n = 100$) učne klikotoke, ki so se izkazali za bolj stabilen vir podatkov od krajših klikotokov. V poskusih smo uporabili različne hitrosti pozabljanja, pri obeh skupinah spletnih uporabnikov pa smo najboljše napovedi dobili pri vrednosti $a = 100$. Pomembnost posameznih dogodkov v učnem klikotoku je izračunana glede na čas, ki je pretekel od dogodka do trenutka gradnje profila, kar imenujemo starost dogodka. Padanje pomembnosti dogodkov z njihovo starostjo pri uporabi hitrosti pozabljanja $a = 100$ je prikazana v tabeli 4.2.

Sliki 4.5 in 4.6 prikazujeta kvaliteto uporabniških profilov namiznih in mobilnih uporabnikov, zgrajenih z algoritmi LastAction, AverageActionF in AverageAction. Na



Slika 4.3

Tridimenzionalni prikaz gibanja povprečne kvalitete profilov namiznih uporabnikov v odvisnosti od starosti profila in dolžine učnega klikotoka. Profili so bili zgrajeni z algoritmom AverageAction. Uporabljeni so bili dolgi učni klikotoki (n = 100).



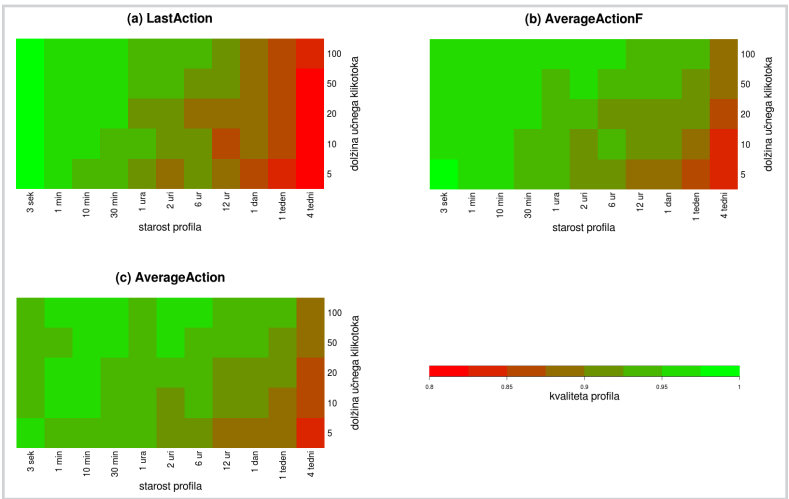
Slika 4.4

Tridimenzionalni prikaz gibanja povprečne kvalitete profilov mobilnih uporabnikov v odvisnosti od starosti profila in dolžine učnega klikotoka. Profili so bili zgrajeni z algoritmom AverageAction. Uporabljeni so bili dolgi učni klikotoki (n = 100).

Tabela 4.2

Vpliv mehanizma časovnega pozabljanja na pomembnost posameznega dogodka glede na njegovo starost. Moč časovnega pozabljanja je določena s parametrom a v enačbi 3.1.

moč časovnega pozabljanja	0 sekund	1 minuta	1 ura
$a = 10$	1.0	0.36	0.22
$a = 100$	1.0	0.53	0.36

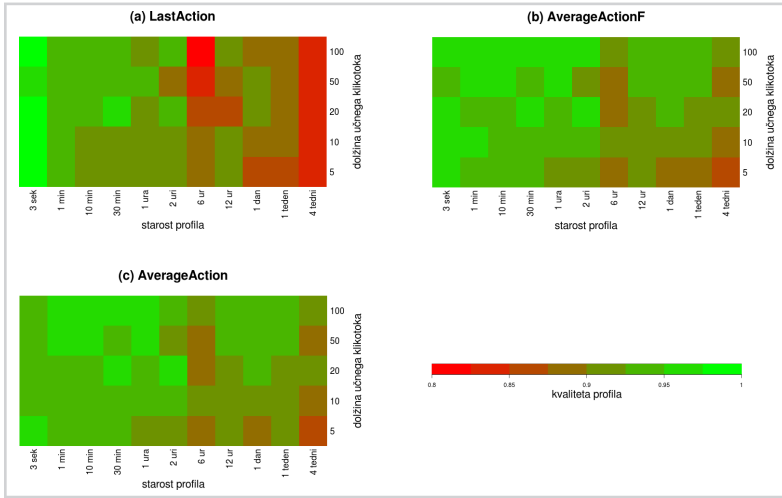


Slika 4.5

Primerjava kvalitete profilov namiznih uporabnikov v odvisnosti od dolžine učnih klikotokov in starosti profilov. Uporabljene so bile tri različne hitrosti pozabljanja in dolgi učni klikotoki ($n = 100$).

sliki 4.5 se dobro vidi, da imata staranje profila in krajšanje učnih klikotokov negativen vpliv na kvaliteto uporabniških profilov. Kratkoročne interese uporabnikov lahko najbolje napovedujemo z uporabo profilov, zgrajenih z algoritmom LastAction, hkrati pa je ta algoritem tudi najmanj primeren za srednje- in dolgoročne napovedi, za katere sta bolj primerna algoritma AverageActionF in AverageAction. Razlika med slednjima je na slikah 4.5 in 4.6 zaradi manjše natančnosti v prikazih kvalitet profilov neopazna.

Na slikah 4.7 in 4.8 je prikazano gibanje kvalitete uporabniških profilov za namizne in mobilne uporabnike pri uporabi dolgih učnih klikotokov in treh različnih hitrosti pozabljanja. Pri obeh skupinah spletnih uporabnikov se je izkazalo, da je smiselno uporabljati hitro pozabljanje za kratkoročne napovedi, zmerno pozabljanje za srednjeročne in izklopljen mehanizem pozabljanja za dolgoročne napovedi. Pri mobilnih upo-



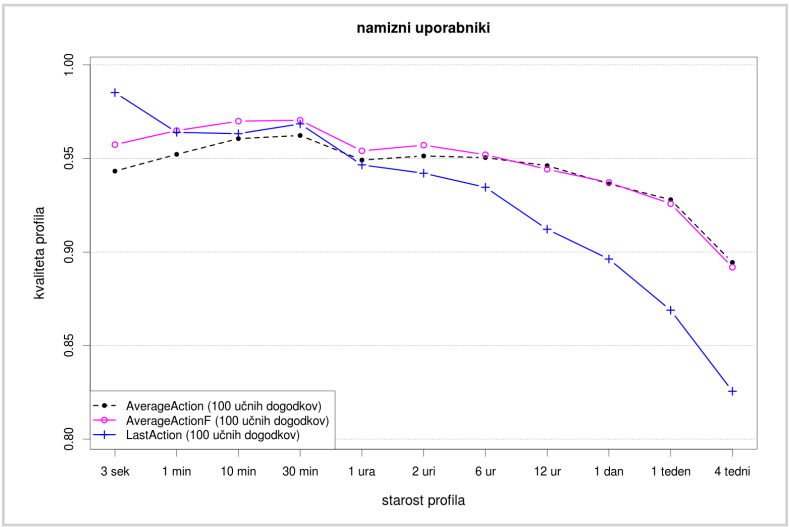
Slika 4.6

Primerjava kvalitete profilov mobilnih uporabnikov v odvisnosti od dolžine učnih klikotokov in starosti profilov. Uporabljeneso bile tri različne hitrosti pozabljanja in dolgi učni klikotoki ($n = 100$).

rabnikih (slika 4.8) je opaziti manjše razlike v kvaliteti profilov, zgrajenih z uporabo zmernega pozabljanja in brez pozabljanja, kot pri namiznih uporabnikih (slika 4.7).

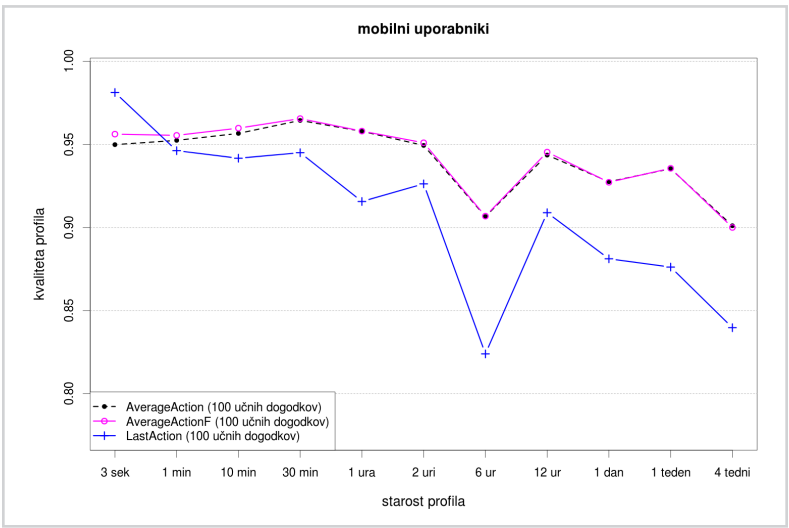
Množica prototipnih profilov je pomemben vir informacij za gradnjo profilov z algoritmom AverageActionFC. Dostopnost do informacij o preteklih aktivnostih uporabnikov oglaševalske mreže nam omogoča, da definiramo prototipne profile na podlagi analize njihovih klikotokov. Iz kalibracijskih podatkov smo izluščili klikotoke posameznih spletnih uporabnikov, za nadaljnjo obdelavo pa smo obdržali le klikotoke z najmanj 100 vnosi. S tem smo se zaščitili pred vplivom velikega števila kratkih klikotokov, ki so lahko posledica:

- spletnih uporabnikov, ki zelo redko zahajajo na spletne strani, ki so del oglaševalske mreže ali
- naprednih spletnih uporabnikov, ki z brisanjem spletnih piškotkov ali uporabo "zasebnostnega" načina delovanja spletnega brskalnika znižujejo učinkovitost sledenja in s tem dajejo vtis, da ima oglaševalska mreža opravka z dosti večjim številom uporabnikov.



Slika 4.7

Primerjava kvalitet profilov namiznih uporabnikov, zgrajenih z uporabo treh različnih hitrosti pozabljanja. Uporabljeni so bili dolgi učni klikotoki ($n = 100$).



Slika 4.8

Primerjava kvalitet profilov mobilnih uporabnikov, zgrajenih z uporabo treh različnih hitrosti pozabljanja. Uporabljeni so bili dolgi učni klikotoki ($n = 100$).

Z algoritmom AverageAction (Algoritem 1) smo zgradili 3403 uporabniških profilov. Pri razvrščanju profilov z algoritmom k-means [22] smo za mero podobnosti uporabili posplošeno kosinusno podobnost (opisana je na strani 23), ki upošteva pri izračunu podobnosti med dvema profiloma tudi strukturo oglaševalske ontologije. Centroide, ki so rezultat razvrščanja s k-means, smo poimenovali prototipni profili in jih v nadaljevanju uporabili za izboljšanje kakovosti uporabniških profilov.

Pri določanju optimalnega števila prototipnih profilov k smo se osredotočili predvsem na dvig kvalitete uporabniških profilov, ki so bili zgrajeni na podlagi kratkih učnih klikotokov, saj smo pri njih opazili veliko slabše srednjeročne in dolgoročne napovedi kot pri dolgih klikotokih. Na kalibracijskih podakih smo tako določili optimalno število prototipov $k = 200$, to vrednost pa smo uspeli kasneje potrditi tudi na testnih podatkih.

Vrednost uteži za popravljanje profilov , v Algoritmu 2 označeno z $w_{correct}$, smo prav tako določili na podlagi poskusov na kalibracijskih podatkih. Ugotovili smo, da lahko kvaliteto profilov, zgrajenih s kratkimi učnimi klikotoki, izboljšamo s katerokoli vrednostjo $0 < w_{correct} < 1$, najboljše rezultate pa smo dobili z uporabo $w_{correct} = 0.9$.

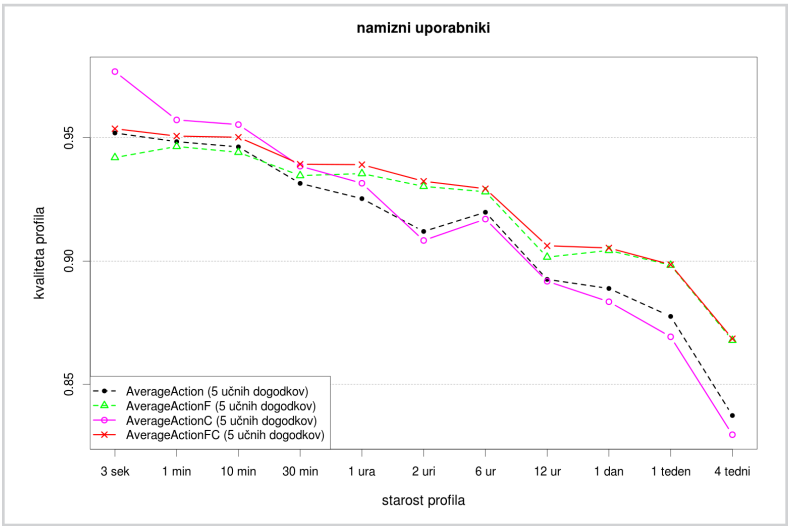
4.2.3 Rezultati

V tem razdelku so predstavljeni rezultati poskusov na podatkih spletne oglaševalske mreže Httpool. Populacijo spletnih uporabnikov oglaševalske mreže smo razdelili na namizne in mobilne uporabnike, rezultati pa so prikazani za vsako od skupin uporabnikov posebej.

Vpliv popravljanja s prototipi na kvaliteto uporabniških profilov

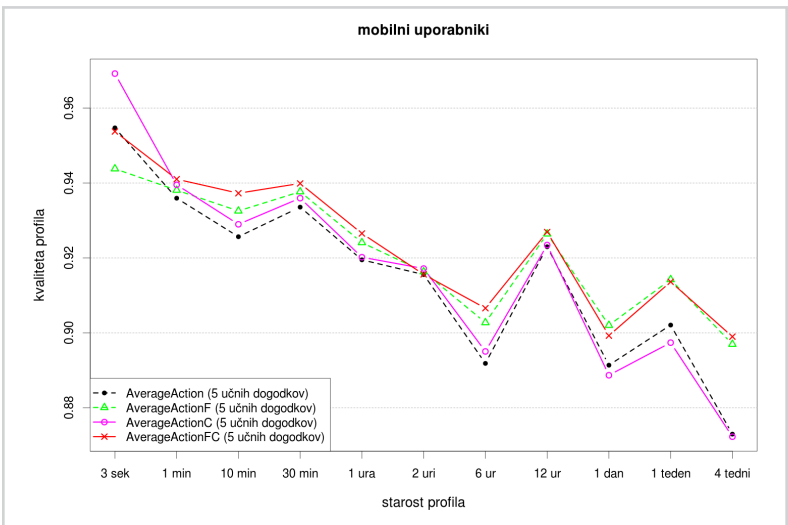
Na slikah 4.9 in 4.10 vidimo, da lahko s popravljanjem s prototipi občutno izboljšamo kvaliteto uporabniških profilov, zgrajenih na podlagi kratkih učnih klikotokov. Algoritem AverageVisitF uporablja časovno pozabljanje, AverageVisitC popravljanje s prototipi, AverageVisitFC pa oboje. V nasprotju s kratkimi so dolgi učni klikotoki očitno veliko boljši vir informacij za gradnjo profilov.

Sliki 4.11 in 4.12 prikazujeta gibanje kvalitete profilov, zgrajenih na učnih klikotokih dolžine 100 dogodkov. Koristnost uporabe popravljanja s prototipi je v tem primeru veliko manjša, ponekod celo neopazna.



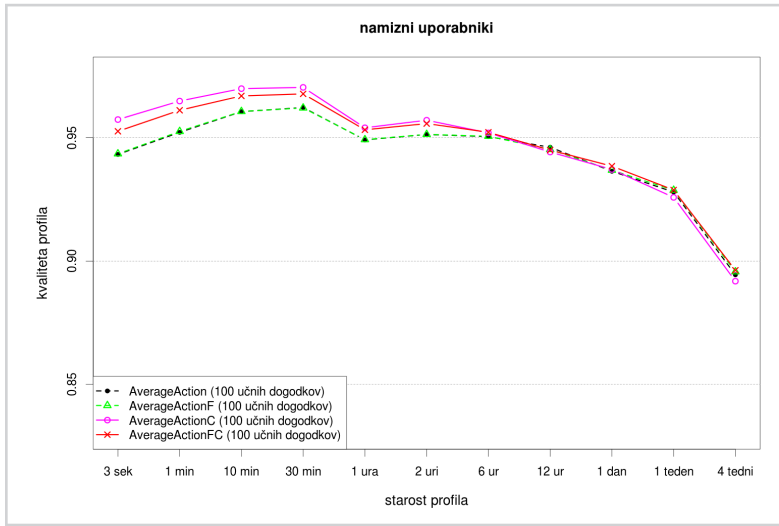
Slika 4.9

Prikazan je vpliv popravljanja profilov s prototipi na kvaliteto profilov namiznih uporabnikov, zgrajenih s kratkimi učnimi klikotoki ($n = 5$).



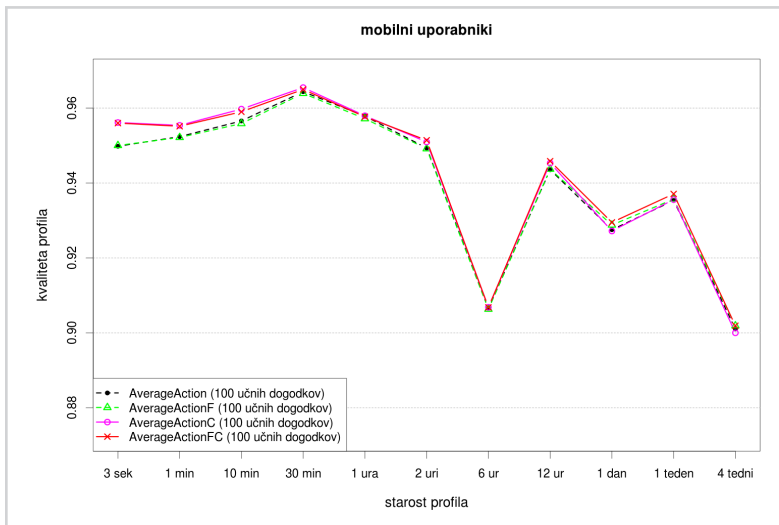
Slika 4.10

Prikazan je vpliv popravljanja profilov s prototipi na kvaliteto profilov mobilnih uporabnikov, zgrajenih s kratkimi učnimi klikotoki ($n = 5$).



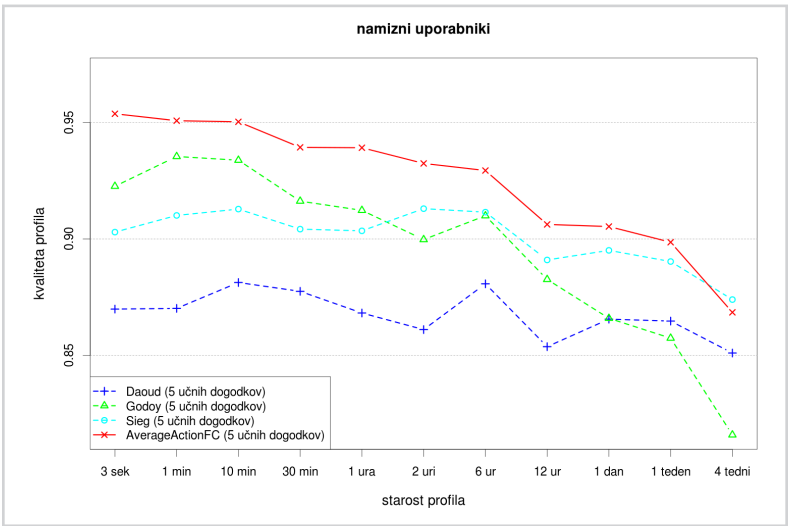
Slika 4.11

Prikazan je vpliv popravljanja profilov s prototipi na kvaliteto profilov namiznih uporabnikov, zgrajenih s dolgimi učnimi klikotoki ($n = 100$).



Slika 4.12

Prikazan je vpliv popravljanja profilov s prototipi na kvaliteto profilov mobilnih uporabnikov, zgrajenih s dolgimi učnimi klikotoki ($n = 100$).



Slika 4.13

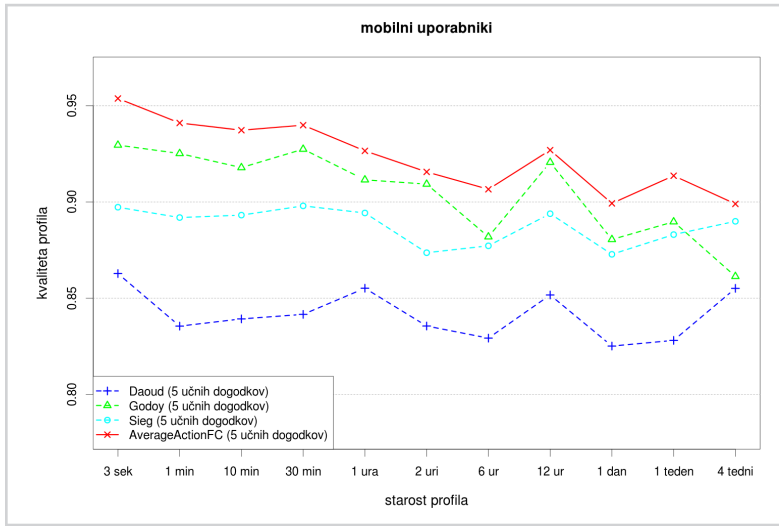
Primerjava metode AverageAction z obstoječimi metodami na populaciji namiznih uporabnikov. Uporabljeni so kratki učni klikotoki dolžine $n = 5$.

Primerjava metode AverageActionFC z obstoječimi metodami

Metodo AverageActionFC smo primerjali s tremi obstoječimi metodami za gradnjo ontoloških profilov uporabnikov: Daoud [18], Godoy [21] in Sieg [8]. Poskusi s kratkimi učnimi klikotoki (sliki 4.13 in 4.14) so pokazali, da lahko z našo metodo zgradimo veliko bolj kvalitetne uporabniške profile kot z obstoječimi metodami.

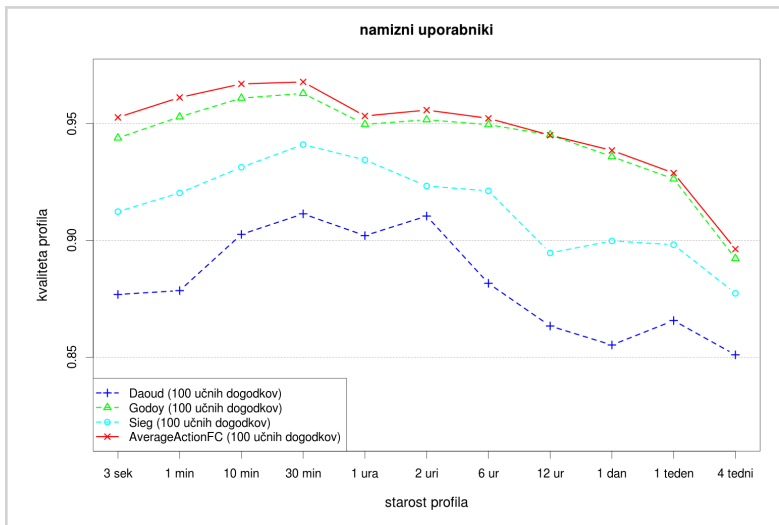
Pri profilih, zgrajenih z dolgimi učnimi klikotokih (sliki 4.15 in 4.16), je prednost metode AverageActionFC pred obstoječimi metodami manjša, a še vedno očitna.

Rezultate smo statistično preverili še z Wilcoxonovim testom predznačenih rangov [25]. Izvedli smo serijo $3 \times 2 \times 11 = 66$ testov, pri čemer smo v vsakem testu primerjali rezultate metode AverageActionFC z eno od treh obstoječih metod, teste pa smo ponovili za obe populaciji spletnih uporabnikov (namizni in mobilni), uporabili pa smo 11 različnih dolžin učnih klikotokov med $5 \leq n \leq 100$. Pri stopnji značilnosti $\alpha = 0.0001$ smo le v enem od testov (primerjava metode AverageActionFC z metodo Godoy na mobilnih uporabnikih, z dolžino učnega klikotoka $n = 20$) dobili vrednost $p > \alpha$. Zaradi večjega števila testov je potrebno pri interpretaciji rezultatov upoštevati še Bonferronijev popravek, kljub temu pa lahko še vedno trdimo, da je naša metoda boljša od obstoječih pri stopnji značilnosti $\alpha = 0.01$.



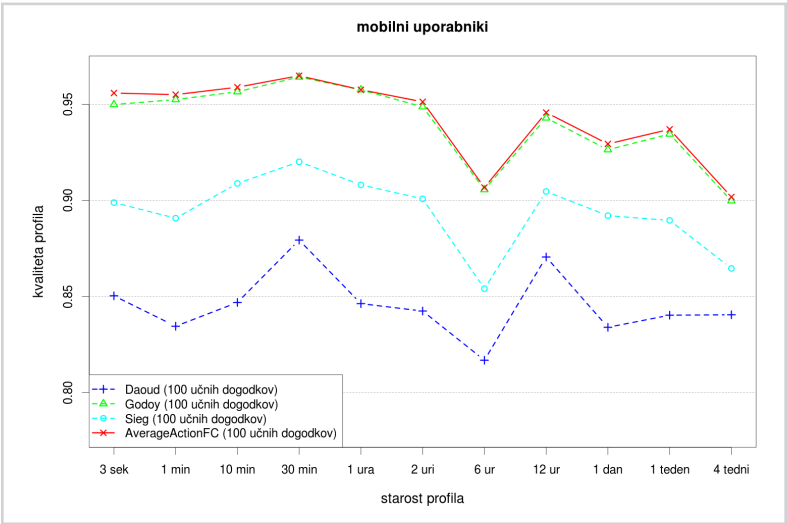
Slika 4.14

Primerjava metode AverageAction z obstoječimi metodami na populaciji mobilnih uporabnikov. Uporabljeni so kratki učni klikotoki dolžine $n = 5$.



Slika 4.15

Primerjava metode AverageAction z obstoječimi metodami na populaciji namiznih uporabnikov. Uporabljeni so dolgi učni klikotoki dolžine $n = 100$.



Slika 4.16

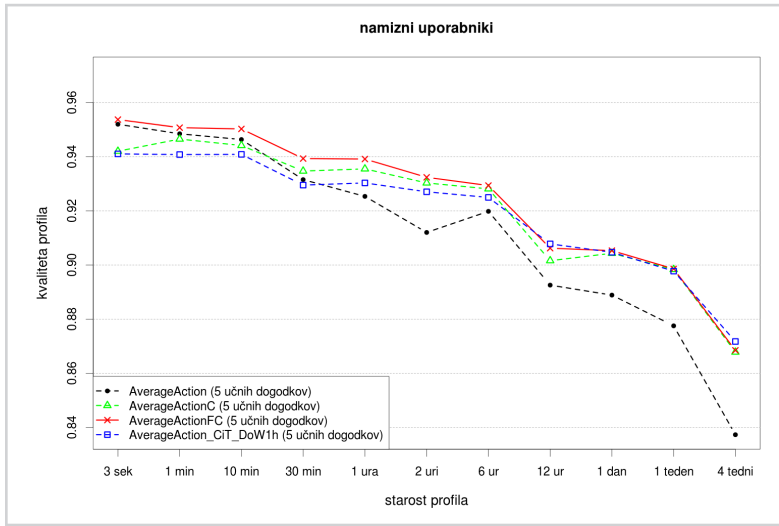
Primerjava metode AverageAction z obstoječimi metodami na populaciji mobilnih uporabnikov. Uporabljeni so dolgi učni klikotoki dolžine $n = 100$.

Metode, ki popravljajo profile s časovnimi statistikami

Med vsemi štirimi različicami časovnih statistik so se najbolje obnesle najbolj podrobne - to so urne statistike za posamezne dneve v tednu. Na slikah 4.17 - 4.18 je ta profilirna metoda označena kot "AverageAction_CiT_DoW1h".

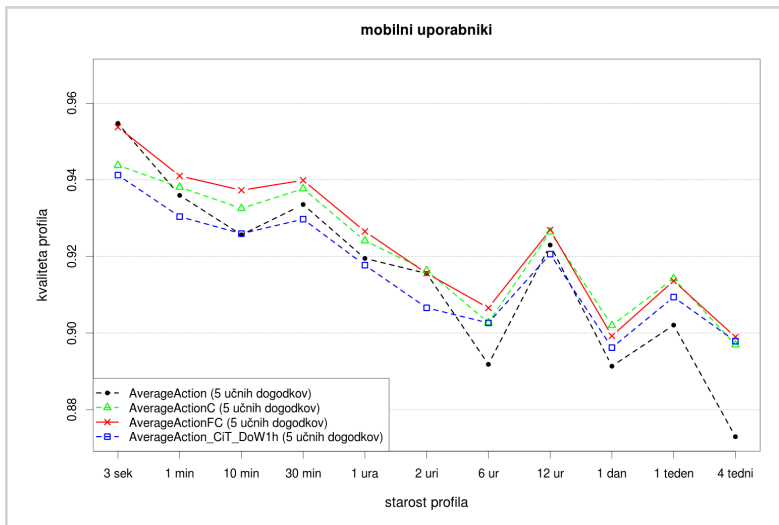
Pri evalvaciji uporabniških profilov, zgrajenih na podlagi kratkih učnih klikotokov (sliki 4.17 in 4.18), se je izkazalo, da ima uporaba časovnih statistik negativen vpliv na kvaliteto napovedi kratkoročnih interesov uporabnikov, pri napovedih dolgoročnih interesov spletnih uporabnikov pa je vpliv izredno pozitiven.

Pri profilih, zgrajenih z dolgimi učnimi klikotoki, je vpliv popravljanja profilov s časovnimi statistikami na kvaliteto uporabniških profilov minimalen (sliki 4.19 in 4.20).



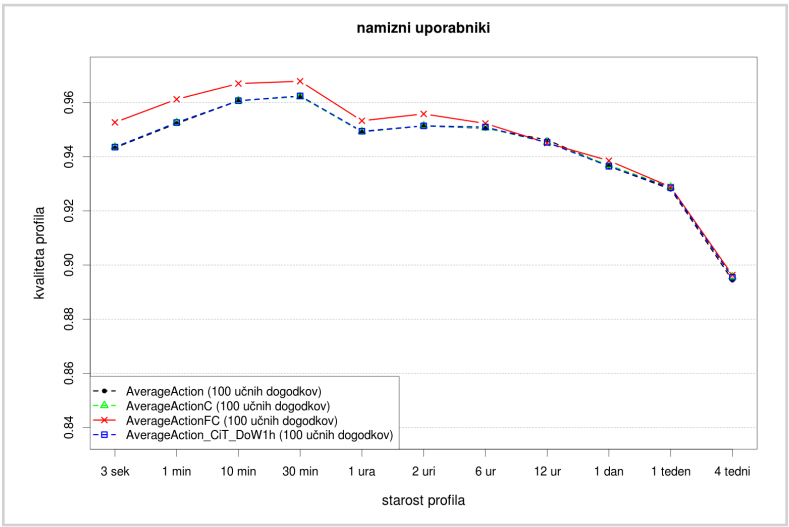
Slika 4.17

Primerjava metode AverageAction in metod, ki uporabljajo popravljanje profilov s prototipi in časovnimi statistikami na populaciji namiznih uporabnikov. Uporabljeni so kratki učni klikotoki dolžine $n = 5$.



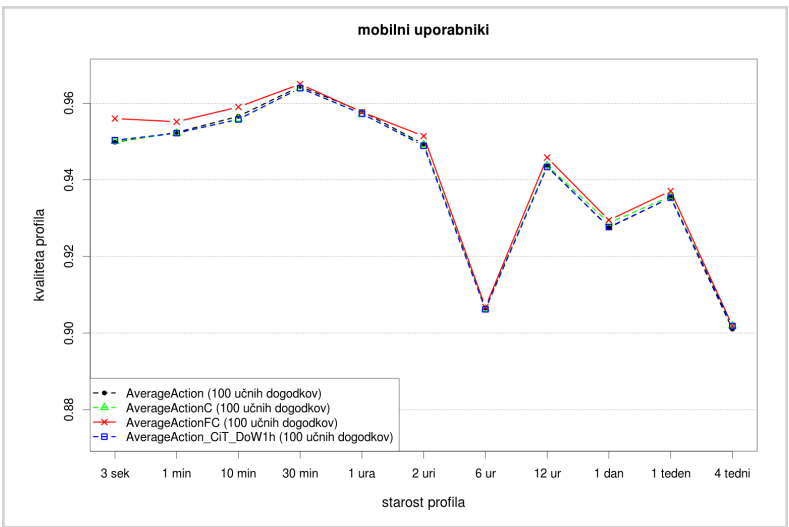
Slika 4.18

Primerjava metode AverageAction in metod, ki uporabljajo popravljanje profilov s prototipi in časovnimi statistikami na populaciji mobilnih uporabnikov. Uporabljeni so kratki učni klikotoki dolžine $n = 5$.



Slika 4.19

Primerjava metode AverageAction in metod, ki uporabljajo popravljanje profilov s prototipi in časovnimi statistikami na populaciji namiznih uporabnikov. Uporabljeni so dolgi učni klikotoki dolžine $n = 100$.



Slika 4.20

Primerjava metode AverageAction in metod, ki uporabljajo popravljanje profilov s prototipi in časovnimi statistikami na populaciji mobilnih uporabnikov. Uporabljeni so dolgi učni klikotoki dolžine $n = 100$.

4.3 Poskusi: Kvaliteta profilov študentov spletne učilnice

Spletna učilnica je osnovana na odprtokodni platformi za e-učenje Moodle [1]. Ta omogoča učiteljem hitro in enostavno objavljanje učnega materiala, seminarskih nalog, obvestil, komunikacijo s študenti preko forumov, izvajanje izpitov, itd. Namen analize obnašanja študentov na spletni učilnici je evalvacija metod za gradnjo profilov in potrditev rezultatov poskusov na domeni oglaševalske mreže.

Pri uporabi te podatkovne množice se moramo zavedati, da obstajajo pri aktivnostih študentov v okviru spletne učilnice znatne vsebinske (ožja računalniška domena in majhen nabor predmetov, ki jih opravljajo v tekočem semestru) in časovne omejitve (npr. roki za oddajo seminarskih nalog, datumi kolokvijev in izpitov). Pri profiliranju spletnih uporabnikov oglaševalske mreže (poglavje 4.2) teh omejitev ni oziroma se izgubijo v množici številnih spletnih strani in uporabnikov.

4.3.1 Opis podatkov spletne učilnice

V strežniške dnevnike sistema Moodle se zapisujejo vse aktivnosti obiskovalcev spletne učilnice. Določeni deli spletne učilnice so na voljo tako prijavljenim kot neprijavljenim obiskovalcem. Med prijavljenimi uporabniki najdemo študente, profesorje, asistente, in administratorje. V okviru tega poskusa smo se omejili na študente, ki predstavljajo večino uporabnikov spletne učilnice.

Pri evalvaciji algoritmov za profiliranje uporabnikov smo se omejili na podatke iz zimskih semestrov, in sicer zato, ker imajo med letnimi semestri študenti več motečih dejavnikov (npr. vmesni izpitni roki za predmete iz zimskih semestrov). Za kalibracijo metod za profiliranje smo uporabili podatke iz zimskega semestra 2011/2012, za evalvacijo in medsebojno primerjavo metod pa podatke iz naslednjega zimskega semestra v študijskem letu 2012/2013. Z uporabo zaporednih zimskih semestrov smo minimizirali razlike v naborih predmetov, ki so na voljo študentom v obeh semestrih.

Vsak vnos v strežniškem dnevniku spletne učilnice vsebuje naslednje podatke o aktivnostih njenih uporabnikov:

- naslov URL obiskane strani,
- čas obiska obiskane strani,

[1] https://moodle.org/

- številka ID prijavljenega uporabnika,
- številka ID predmeta, ki mu pripada obiskana stran,
- ime dela (modula) sistema Moodle, ki je bil uporabljen in
- akcija, ki jo je izvedel uporabnik.

Pred uporabo dnevnikov za gradnjo in evalvacijo profilov smo izločili uporabniške akcije, kot so prijava in odjava iz učilnice, urejanje uporabnikovih lastnih podatkov ipd.

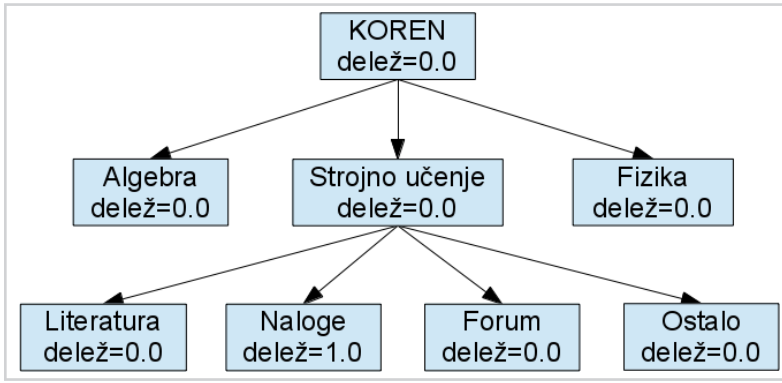
Uporabljena ontologija. Za gradnjo ontoloških profilov študentov smo zgradili namensko ontologijo z drevesno strukturo (glej sliko 4.21). Drevo ima tri nivoje:

1. koren drevesa,
2. vozlišča na drugem nivoju predstavljajo posamezne študijske programe in predmete, ki so na voljo v spletni učilnici,
3. za vsako vozlišče na drugem nivoju so na tretjem nivoju definirana štiri vozlišča, ki predstavljajo uporabnikove akcije v okviru drugonivojskega vozlišča
   - branje *literature*, ki je v okviru predmeta na voljo študentu,
   - ogledi in reševanje *nalog* (sem spadajo seminarske naloge, ankete, kvizi in izpiti)
   - branje in sodelovanje na *forumu* ter
   - *ostalo*.

Kot je prikazano na sliki 4.21, smo vsako uporabnikovo akcijo klasificirali v eno od vozlišč na tretjem nivoju. Uporabnikova akcija je točkovana tako, da je primernemu vozlišču na tretjem nivoju pripisana utež $w = 1$, vsem ostalim pa $w = 0$.

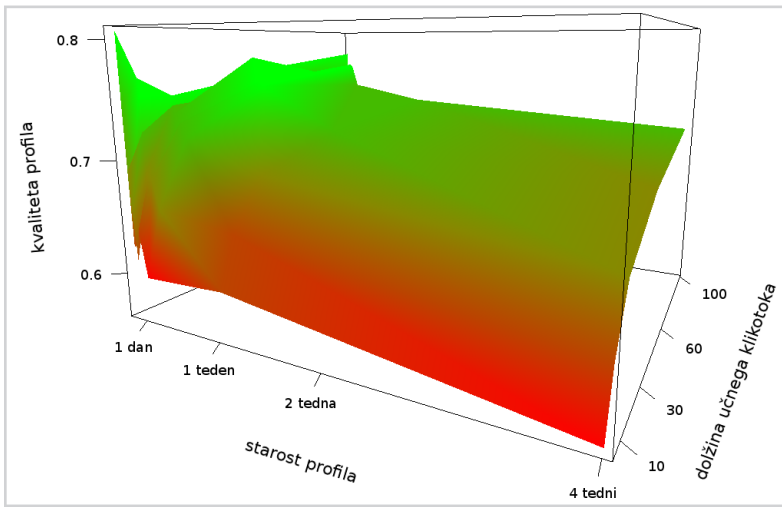
4.3.2 Kalibracija metode AverageActionFC

Moč časovnega pozabljanja , ki jo uravnavamo prek parametra a v enačbi 3.1, smo prilagodili srednje- in dolgoročnim napovedim interesov študentov v spletni učilnici. Na sliki 4.22 je prikazano gibanje kvalitete profilov študentov, zgrajenih z algoritmom



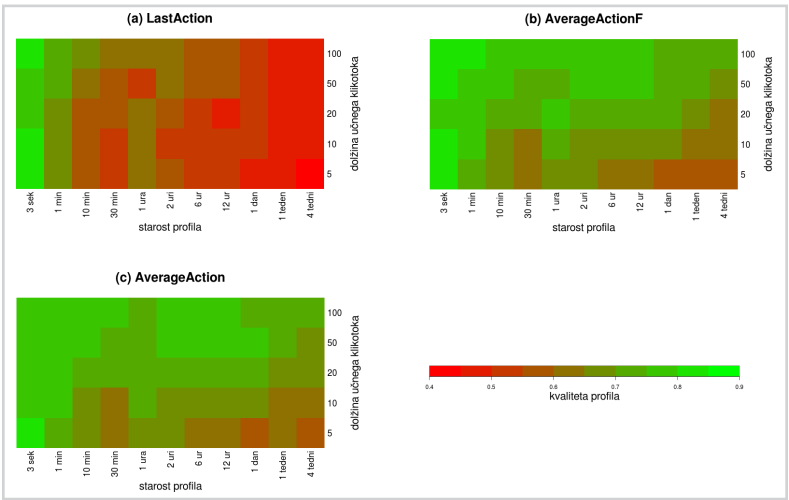
Slika 4.21

Na sliki je prikazan manjši izsek ontologije, ki je bila zgrajena za profiliranje študentov spletne učilnice. Vozlišča na drugem nivoju predstavljajo predmete, vozlišča na tretjem nivoju pa možne uporabnikove akcije. Prikazan je primer reševanja naloge v okviru predmeta Strojno učenje.



Slika 4.22

Tridimenzionalni prikaz gibanja povprečne kvalitete profilov študentov v odvisnosti od starosti profila in dolžine učnega klikotoka. Profili so bili zgrajeni z algoritmom AverageAction.



Slika 4.23

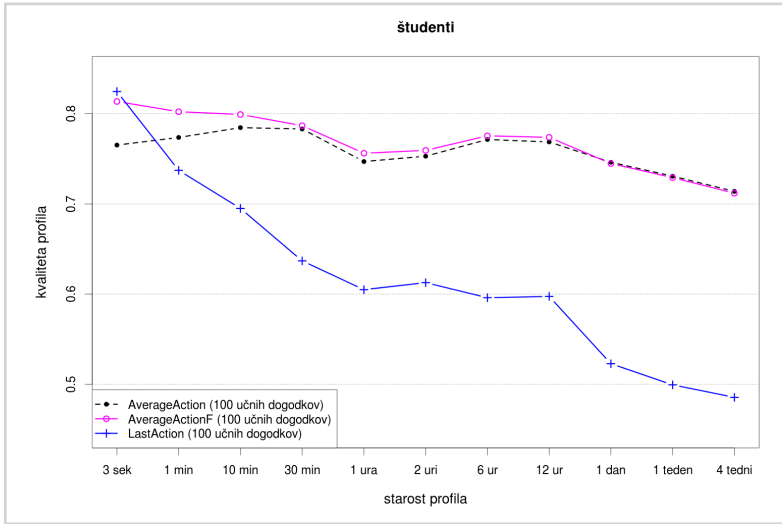
Primerjava kvalitete profilov študentov v odvisnosti od dolžine učnih klikotokov in starosti profilov. Uporabljeneso bile tri različne hitrosti pozabljanja in dolgi učni klikotoki ($n = 100$).

AverageAction, se pravi brez časovnega pozabljanja. Opaziti je podobno obliko, kot na slikah 4.3 in 4.4, le da je padec kvalitete profilov s časom manj očiten.

Pri kalibraciji smo ugotovili, da je pri tej podatkovni množici potrebna večja moč časovnega pozabljanja $a = 10$ (glej tabelo 4.2). Pri tej vrednosti se pomembnost posameznih dogodkov znižuje hitreje kot pri profiliranju spletnih uporabnikov v oglaševalski mreži.

Vpliv dolžine učnih klikotokov in starosti na kvaliteto profilov je dobro viden na sliki 4.23. Podobno, kot na slikah 4.5 in 4.6, je tudi s slike 4.23 očitno, da lahko z daljšimi učnimi klikotoki zgradimo bolj kvalitetne profile, s staranjem profilov pa njihova kvaliteta pada. Algoritem LastAction je najbolj primeren za modeliranje uporabnikovih kratkoročnih interesov, algoritma AverageActionF in AverageAction pa za srednje- in dolgoročne interese.

Na sliki 4.24 vidimo vpliv časovnega pozabljanja na kvaliteto profilov študentov v spletni učilnici. Pri študentih je za razliko od spletnih uporabnikov oglaševalske mreže prednost algoritma LastAction (absolutno pozabljanje) pri kratkoročnih napovedih veliko manj izrazita.



Slika 4.24

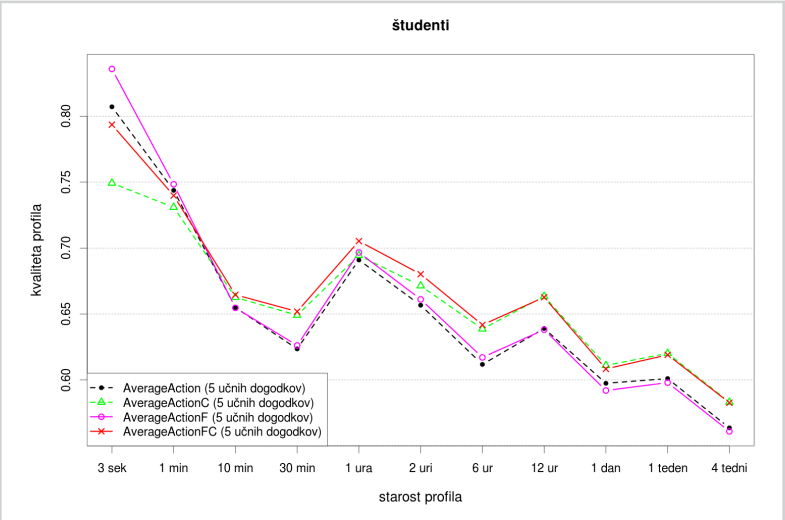
Primerjava kvalitet profilov študentov, zgrajenih z uporabo treh različnih hitrosti pozabljanja.

Množico prototipnih profilov smo definirali na podlagi poskusov na kalibracijskih podatkih, to je podatkih iz zimskega semestra 2011/2012. Profile študentov, ki smo jih zgradli s preprostim algoritmom AverageAction (glej Algoritem 1), smo z algoritmom k-means razvrstili v gruče, dobljene centroide pa smo uporabili za popravljanje profilov.

V skladu s številom študijskih programov smo med kalibracijo omejili iskanje optimalnega števila prototipov na $k \leq 20$. Poskusi na kalibracijskih podatkih so pokazali, da lahko zgradimo najbolj kvalitetne profile z uporabo le dveh prototipnih profilov. Ta vrednost se je pokazala za optimalno tudi z obsežnejšimi poskusi na podatkih za evalvacijo in primerjavo metod. Nizko število nudi zanimiv vpogled v razlike med obema prototipnima profiloma študentov:

- *Prototip A* (pridni študenti) predstavlja študente, ki pogosto obiskujejo strani posameznih predmetov, berejo literaturo, sodelujejo na forumih in rešujejo razne naloge v okviru spletne učilnice.

- *Prototip B* (manj aktivni študenti) povzema aktivnosti študentov, ki so obiskovali predvsem uvodne strani študijskih programov in uporabljali storitve, ki niso



Slika 4.25

Prikazan je vpliv popravljanja profilov s prototipi na kvaliteto študentskih profilov, zgrajenih s kratkimi učnimi klikotoki ($n = 5$).

neposredno povezane s predmeti, recimo pošiljanje sporočil drugim uporabnikom.

Pri kalibraciji uteži za popravljanje profilov s prototipi se je izkazalo, da je optimalna vrednost $w_{correct} = 0.7$, kar je nekoliko nižja vrednost kot pri profiliranju uporabnikov oglaševalske mreže.

4.3.3 Rezultati

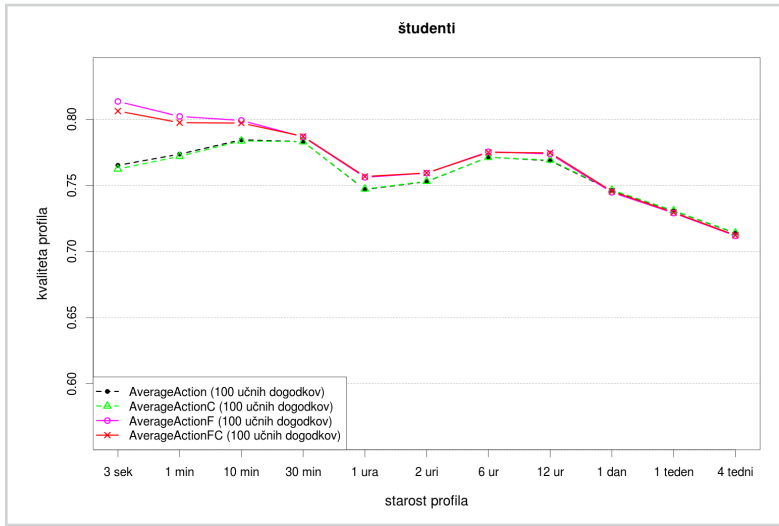
Vpliv popravljanja s prototipi na kvaliteto uporabniških profilov

Na sliki 4.25 vidimo, da popravljanje profilov s prototipi občutno izboljša kvaliteto profilov, zgrajenih na podlagi kratkih učnih klikotokov.

Pri dolgih učnih klikotokih (slika 4.26) se je izkazalo, da uporaba popravljanja profilov s prototipi ni smiselna, saj nima večjega vpliva na kvaliteto profilov.

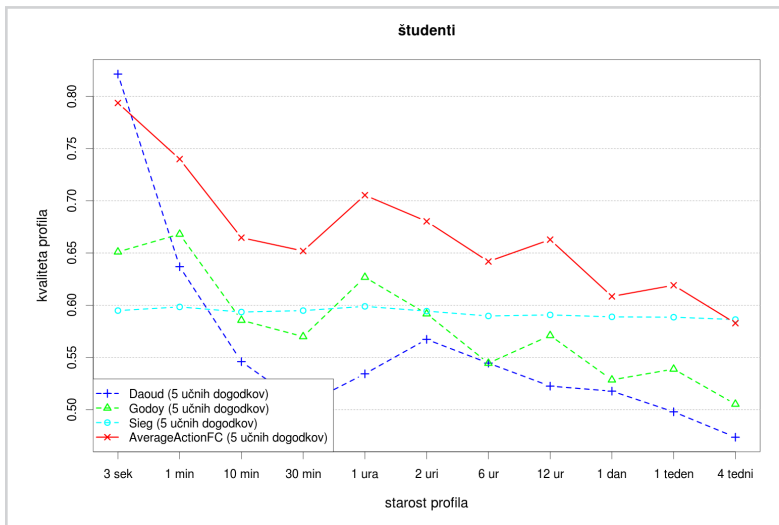
Primerjava metode AverageActionFC z obstoječimi metodami

Metodo AverageActionFC smo primerjali z obstoječimi metodami avtorjev Daoud [18], Godoy [21] in Sieg [8]. Pri uporabi kratkih (slika 4.27) in dolgih učnih klikotokov (slika 4.28) se je metoda AverageActionFC izkazala za najboljšo.



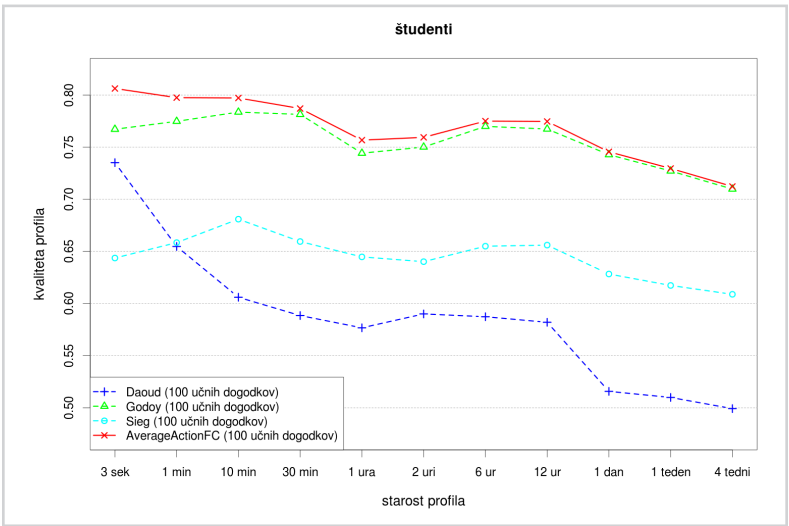
Slika 4.26

Prikazan je vpliv popravljanja profilov s prototipi na kvaliteto študentskih profilov, zgrajenih s dolgimi učnimi klikotoki ($n = 100$).



Slika 4.27

Primerjava metode AverageActionFC z obstoječimi metodami. Študentski profili so zgrajeni s kratkimi učnimi klikotoki ($n = 5$).



Slika 4.28

Primerjava metode AverageActionFC z obstoječimi metodami. Študentski profili so zgrajeni z dolgimi učnimi klikotoki ($n = 100$).

Primerjavo profilirnih metod smo okrepili še z Wilcoxonovim testom predznačenih rangov [25]. V seriji $3 \times 5 = 15$ testov smo metodo AverageActionFC primerjali z vsako od obstoječih metod, uporabili pa smo 5 različnih dolžin učnih klikotokov dolžin $5 \leq n \leq 100$ dogodkov. Z uporabo Bonferronijevega popravka in stopnje značilnosti $\alpha = 0.0001$ smo brez težav potrdili, da je metoda AverageActionFC boljša od ostalih. Vse p-vrednosti so bile namreč manjše od 10^{-23}.

4.4 Ugotovitve

Profiliranje uporabnikov v oglaševalski mreži in študentov v spletni učilnici se je izkazalo za težavno. V poskusih na obeh podatkovnih množicah se je že algoritem AverageAction, ki izračuna utež vsakega koncepta v ontologiji kot preprosto povprečje vseh istoležečih uteži v dogodkih iz uporabnikovega učnega klikotoka, izkazal za nepričakovano dobrega. Marsikateri od razvitih algoritmov za gradnjo ontoloških profilov, ki ni opisan v disertaciji, se je namreč izkazal za slabšega. Vzrok za to je verjetno zelo nepredvidljivo obnašanje uporabnikov.

Preskok v kvaliteti se je zgodil šele, ko smo poleg uporabnikovega učnega klikotoka za gradnjo profila uporabili še dodatne vire informacij - prototipne profile in časovne

statistike. Popravljanje uporabniških profilov s prototipnimi profili občutno izboljša srednje- in dolgoročne interese uporabnikov, predvsem takrat, ko imamo za gradnjo profila na voljo le malo podatkov.

Uporaba profilov spletnih uporabnikov v sistemih za priporočanje

5

5.1 Pregled področja

Sistemi za priporočanje so mlado in izredno aktivno raziskovalno področje. Razloge za to lahko iščemo v razmahu svetovnega spleta in premiku storitev (npr. trgovina in oglaševanje) na spletne platforme. Spletna podjetja so kmalu spoznala, da je potrebno uporabnikom v omejenem času, ki ga preživijo na njihovih spletnih straneh, ponuditi čim bolj relevantne in zanimive vsebine, saj lahko na ta način izboljšajo uporabniško izkušnjo in/ali povečajo prodajo. Med spletnimi aplikacijami, ki temeljijo na uporabi sistemov za priporočanje, najdemo:

- spletne trgovine (npr. Amazon.com[1], eBay[2]),
- spletno oglaševanje (npr. Google AdWords[3]),
- multimedijske portale (npr. Netflix[4], Hulu[5], Pandora[6]) in
- personalizirane spletne iskalnike (npr. Google[7], Yahoo![8]).

Naloga sistema za priporočanje [27] je ocenjevanje, ali in koliko bo določena vsebina všeč danemu uporabniku [28]. Ta problem je običajno predstavljen kot problem *napovedovanja uporabnikovih ocen* za dane vsebine, poenostavljena različica tega problema pa je *rangiranje vsebin* glede na všečnost uporabniku [29].

Sisteme za priporočanje delimo na:

- sisteme z *vsebinskim izbiranjem*,
- sisteme z *izbiranjem na podlagi sodelovanja* in
- *hibridne* sisteme.

[1] http://www.amazon.com/
[2] http://www.ebay.com/
[3] https://www.google.si/adwords/
[4] https://www.netflix.com/
[5] http://www.hulu.com/
[6] http://www.pandora.com/
[7] https://www.google.com/
[8] https://www.yahoo.com/

5.1.1 Vsebinsko izbiranje

Vsebinsko izbiranje temelji na uporabi semantičnih informacij o vseh udeleženih entitetah (uporabnikih in vsebinah).

Običajno so vsi uporabniki in vsebine predstavljeni z množico atributov [7], npr. spol, starost in tematika, seznam priporočil za uporabnika pa se zgradi na podlagi podobnosti med uporabnikovim profilom u in razpoložljivimi vsebinami i (enačba 5.1). Za napovedovanje ocene $score_{u,i}$ uporabnika u za vsebino i se največkrat uporabljata Pearsonov korelacijski koeficient in kosinusna podobnost. Na podlagi izračunov uporabnikovih ocen za vse razpoložljive vsebine lahko omejimo priporočila le na vsebine, ki se najbolj skladajo z uporabnikovim profilom.

$$score_{u,i} = sim(u, i) \tag{5.1}$$

Gradnja uporabnikovega profila (profiliranje) temelji na analizi uporabnikovih preteklih interakcij z vsebinami. Poseben primer sistema za priporočanje, ki gradi profil le na osnovi zadnje uporabnikove interakcije s sistemom, imenujemo v spletnem oglaševanju *kontekstualni sistem za priporočanje*.

Prednosti vsebinskega izbiranja:

- preprosta razlaga priporočil,
- pri novih vsebinah ne trpi za problemom hladnega zagona,
- hitrost sistema za priporočanje je neodvisna od števila uporabnikov in
- sistem ni dovzeten na morebitne zlonamerne napade uporabnikov.

Slabosti vsebinskega izbiranja:

- za delovanje sta potrebna profiliranje uporabnikov in semantična analiza vsebin,
- profili novih uporabnikov so lahko vprašljive kvalitete,
- sistem lahko uporabniku vedno znova priporoča iste vsebine (majhna raznolikost priporočil) in
- hitrost sistema za priporočanje je odvisna od števila možnih vsebin.

Najbolj znana komercialna sistema za priporočanje z vsebinskim izbiranjem sta spletni radio Pandora[9] in spletni vmesnik do filmske podatkovne baze IMDB[10].

5.1.2 Izbiranje na podlagi sodelovanja

V zadnjih letih so raziskave pokazale, da dajejo sistemi z izbiranjem na podlagi sodelovanja boljša priporočila kot vsebinsko izbiranje. Ti sistemi za delovanje ne potrebujejo semantičnih informacij o vsebinah ali uporabnikih, temveč uporabljajo le podatke o interakcijah uporabnikov s sistemom. Največkrat se uporabljajo le eksplicitne povratne informacije, kot je na primer uporabnikova ocena ogledane vsebine. Uporabniku se predlagajo vsebine glede na pretekle interakcije njega in njemu podobnih uporabnikov s sistemom. Podobnost med uporabniki je tu definirana na podlagi interakcij uporabnikov s sistemom in ne na podlagi njihovih semantičnih profilov.

Definicija 7: Matrika povratnih informacij (angl. feedback matrix) predstavlja akumulacijo interakcij uporabnikov s sistemom za priporočanje v določenem obdobju. Vrstice predstavljajo posamezne uporabnike, stolpci pa vsebine. Vsebina celice R_{ui} tako nudi vpogled v interakcije uporabnika u z vsebino i. Najpogosteje se uporablja matrika, v kateri vrednost celice predstavlja informacijo, ali je uporabnik v določenem obdobju kliknil na oglas, ali ne (1 ali 0).

Za izbiranje s sodelovanjem *na osnovi pomnjenja* je značilno, da držijo "v spominu" vse opravljene interakcije uporabnikov s sistemom za priporočanje [28]. Ko iščejo primerne vsebine za določenega uporabnika, se največkrat uporablja ena od naslednjih strategij:

1. Priporoči vsebine, ki so podobne vsebinam, ki jih je ta uporabnik pozitivno ocenil. Na ta način deluje npr. spletni radio Last.fm[11].

2. Priporoči vsebine, ki so jih pozitivno ocenili njemu najbolj podobni uporabniki. Najbolj znan komercialni sistem za priporočanje te vrste je Amazon.com[12]. Ta

[9] http://www.pandora.com/
[10] http://www.imdb.com/
[11] http://www.last.fm/
[12] http://www.amazon.com/

danemu uporabniku priporoča izdelke, ki so jih kupili uporabniki, ki so v preteklosti že kupili enake izdelke kot dani uporabnik - "users who bought X also bought Y".

Sistemi na osnovi pomnjenja imajo lahko poleg ostalih težav pri izbiranju s sodelovanjem še probleme z velikimi količinami informacij (nizka stopnja skalabilnosti) in redkimi matrikami povratnih informacij.

Pri *izbiranju s sodelovanjem na osnovi modela* se pretekle interakcije uporabnikov s sistemom ne hranijo v spominu, temveč se na podlagi njih zgradi kompaktnejši model. Ti sistemi so manj občutljivi na redke povratne informacije, po drugi strani pa se lahko pri učenju modela izgubijo pomembni detajli. V raziskavah se za gradnjo sistemov za priporočanje z izbiranjem na podlagi sodelovanja uporabljajo najrazličnejše metode podatkovnega rudarjenja. Pazzani [30] je ocenjeval verjetnost, da bo uporabnik pozitivno ocenil dokument z Bayesovim klasifikatorjem.

Za najučinkovitejše sisteme z izbiranjem na podlagi sodelovanja veljajo sistemi, ki uporabljajo *matrično faktorizacijo* [31, 32]. Osnovni princip teh sistemov je aproksimacija ogromne matrike povratnih informacij s produktom dveh ali več manjših matrik. Matrika povratnih informacij je običajno redka matrika, v katerih je velika večina celic prazna. Običajno se delež praznih celic giblje med 95% in 99%. Nekateri postopki matrične faktorizacije so prilagojeni delu z redkimi matrikami in so pri tem veliko bolj učinkoviti.

Velik problem pri izbiranju s sodelovanjem je t.i. *problem hladnega zagona*. Za svoje delovanje potrebujejo namreč sistemi na podlagi izbiranja s sodelovanjem veliko količino informacij o interakcijah uporabnikov s sistemom, do problema pa pride, ko se v sistemu za priporočanje pojavi nova entiteta:

- *nov uporabnik*, ki ni podal še nobene ocene ali
- *nova vsebina*, ki jo je potrebno dodati med priporočila in za katero nimamo podane nobene ocene s strani uporabnikov.

V spletnem okolju je največji problem stalen dotok novih uporabnikov, za katere je potrebno kljub pomanjkanju informacij sestaviti sezname priporočil. Za reševanje problema hladnega zagona se običajno uporabljajo druge razpoložljive informacije o uporabnikih, npr. spol, starost, lokacija in prijatelji v socialnem omrežju [33]. V

socialnih omrežjih je problem hladnega zagona zelo lahko rešljiv, saj novi uporabniki hitro "napolnijo" svoj profil z uporabnimi informacijami.

Zunaj spletnih omrežij je problem hladnega zagona veliko težje rešljiv. Sistemi za priporočanje ga rešujejo na različne načine, odvisno od razpoložljivih informacij. Park in sod. [34] so s t.i. filterboti vnesli v sistem za priporočanje domensko znanje v obliki umetnih uporabnikov s predefiniranimi vrednostmi atributov. Gantner in sod. [35] so v svojem pristopu problem hladnega zagona reševali z uporabo kontekstualnih informacij. S sistemom linearnih enačb so preslikali vrednosti kontekstualnih atributov v vrednosti faktorjev v faktorskih matrikah. To preslikavo so uspešno uporabili za napovedovanje vrednosti faktorjev novih uporabnikov. Na soroden način so se reševanja hladnega zagona lotili Nguyen in sod. [36], ki pa so za napovedovanje faktorjev uporabili pravila. V mobilnih sistemih pa se za priporočanje največkrat uporabljajo podatki o trenutni lokacija uporabnika [37].

Prednosti izbiranja s sodelovanjem na osnovi modela:

- priporočila so raznolika in odsevajo popularnost vsebin in
- sistemi nimajo performančnih problemov pri velikem številu uporabnikov ali vsebin.

Slabosti izbiranja s sodelovanjem na osnovi modela:

- razlaga priporočil ni preprosta,
- za gradnjo dobrega modela potrebuje veliko količino preteklih interakcij uporabnikov s sistemom,
- gradnja modela je lahko časovno zahtevna, poleg tega se lahko pri učenju izgubijo pomembne informacije,
- trpi za problemom hladnega zagona (novi uporabniki ali nove vsebine) in
- sistemi so občutljivi na zlonamerne napade uporabnikov, usmerjene na delovanje sistema.

5.1.3 Hibridni sistemi za priporočanje

Hibridni sistemi za priporočanje uporabljajo mešanico vsebinskega izbiranja in izbiranja s sodelovanjem in so lahko implementirani na različne načine. Razvoj in implementacija hibridnih sistemov zahtevata veliko dodatnega napora, po kvaliteti priporočil pa veljajo za "state-of-the-art". S hibridnimi pristopi se lahko rešimo marsikaterih slabosti, ki tarejo osnovne sisteme, npr. ponavljajočih se priporočil pri vsebinskem izbiranju ali hladnega zagona pri izbiranju s sodelovanjem.

Kvaliteto priporočil lahko zvišamo na primer z uporabo kontekstualnih [38, 39] ali lokacijskih informacij [37] že med postopkom matrične faktorizacije. S kompleksnejšimi pristopi se po drugi strani vedno bolj otežuje razlaga priporočil, ki je pomembna predvsem zaradi zagotavljanja transparentnosti pri delovanju sistemov za priporočanje, kar ima lahko velik vpliv na zaupanje uporabnikov [40].

Večkrat se je izkazalo, da je razlika med naivnim predlaganjem (npr. nepersonaliziranim predlaganjem najpopularnejših vsebin) in najboljšimi pristopi relativno majhna, kar kaže na težavnost problema priporočanja [32]. Na tekmovanju "Netflix Prize" je leta 2009 zmagalo moštvo, sestavljeno iz treh raziskovalnih skupin z imenom "BellKor's Pragmatic Chaos". Za napovedovanje uporabniških ocen vsebin so razvili več kot 100 sistemov za priporočanje. Za končno napoved ocene so uporabili t.i. gradientno pospešena odločitvena drevesa (angl. gradient boosted decision trees), s katerimi so združili napovedi večjega števila napovednih modelov.

5.2 Podatkovna množica

Za evalvacijo sistemov za priporočanje smo uporabili le podatke spletne oglaševalske mreže Httpool. Podatki spletne učilnice, ki smo jih uporabljali v prejšnjem poglavju za profiliranje študentov, niso primerni za evalvacijo sistemov za priporočanje, saj ne vsebujejo nobenih povratnih informacij uporabnikov o prebranih vsebinah.

Strežniški dnevniki oglaševalske mreže vsebujejo podatke o interakcijah spletnih uporabnikov z oglaševalskim sistemom. Za nas najpomembnejši tipi uporabniških interakcij so:

- ogled spletne strani,
- ogled oglasa in
- klik na oglas.

V poskusih uporabljamo dvomesečni izsek strežniških dnevnikov (oktober in november 2011), omejili pa smo se le na uporabnike, ki so v tem obdobju kliknili na vsaj en oglas.

Za prikazovanje oglasov na spletnih straneh se je v omenjenem obdobju uporabljalo t.i. kontekstualno oglaševanje. To pomeni, da je spletni strežnik izbiral primerne oglase za prikaz na podlagi semantične podobnosti med oglasom in vsebino spletne strani. V kolikor se je na spletni strani nahajalo več oglasnih blokov, se je izbor oglasov prilagodil tako, da so se v prvem bloku prikazali glede na kontekstualno ujemanje "najboljši" oglasi, v drugem malo slabši itn. Vrstni red oglasnih blokov je določen glede na njihovo lokacijo v izvorni kodi HTML.

1. *Učni podatki* za sisteme za priporočanje (1 mesec). V tem obdobju smo za sisteme z vsebinskim izbiranjem iz klikotokov (angl. clickstream) uporabnikov na podlagi obiskov spletnih strani gradili njihove profile, za sisteme na osnovi matrične faktorizacije pa na podlagi ogledov in klikov na oglase gradili matriko povratnih informacij (angl. feedback matrix).

2. *Podatki za evalvacijo* sistemov za priporočanje (1 mesec). Vsak klik na oglas, ki se je zgodil v tem obdobju, smo v poskusih uporabili za evalvacijo sistemov za priporočanje. Z vsakim sistemom za priporočanje samo za spletnega uporabnika najprej zgradili seznam priporočil. Na podlagi seznama priporočil in oglasa, ki je bil v resnici kliknjen, smo ocenili uspešnost posameznih sistemov za priporočanje.

Množica aktivnih oglasov

Za trenutek, ko je do sistema za priporočanje prišel zahtevek za seznam predlaganih oglasov, smo z uporabo oglaševalske podatkovne baze zgradili množico aktivnih oglasov. V tej množici se nahajajo vsi oglasi, ki so bili v tistem trenutku del aktivne oglaševalske kampanje. Tako sestavljena množica aktivnih oglasov (število oglasov se je gibalo med 443 in 567) je "optimistična", saj lahko vključuje tudi oglase, ki iz različnih vzrokov v tistem trenutku niso mogli biti prikazani na spletni strani:

1. Oglas je bil spletnemu uporabniku *že večkrat prikazan* in bil zato v procesu izbire primernih oglasov začasno izločen.

2. Oglas ni mogel biti prikazan zaradi *omejitev dnevne porabe sredstev* ali *omejitev skupne porabe sredstev* na nivoju oglaševalske skupine ali oglaševalske kampanje.

3. Oglas ni smel biti prikazan zaradi želje založnika - z nastavitvami je na primer omogočeno, da se na spletni strani ne kažejo oglasi za storitve konkurenčnih podjetij.

5.3 Metodologija testiranja

5.3.1 Osnovni sistemi za priporočanje

Za potrebe raziskovalnega dela smo razvili večje število sistemov za priporočanje.

Naključno priporočanje vsebin uredi vsebine v naključen vrstni red in poda seznam priporočil uporabniku.

Sistemi z *vsebinskim izbiranjem* predlagajo vsebine glede na podobnost med uporabnikovim profilom in semantičnimi informacijami o vsebinah. Za mero podobnosti uporabljamo posplošeno kosinusno podobnost, za gradnjo profilov pa naslednje metode: LastAction, AverageAction, AverageActionFC, Daoud [18], Godoy [21] in Sieg [8]. Vsi razviti sistemi z *izbiranjem na osnovi sodelovanja* temeljijo na matrični faktorizaciji (več o tem v razdelku 5.3.1).

Matriko povratnih informacij smo zgradili s standardnim pristopom, pri katerem z vsebino celice R_{ui} opišemo interakcije uporabnika u z oglasom i v določenem časovnem obdobju. V kolikor je uporabnik v tem obdobju oglas videl in nanj tudi kliknil, smo v celico vpisali vrednost $R_{ui} = 1$. Če je uporabnik oglas videl, a nanj ni kliknil, smo v celico vpisali vrednost $R_{ui} = 0$. V primeru, da uporabnik oglasa v tem obdobju ni videl niti enkrat, ostane celica prazna.

Z matrično faktorizacijo (glej enačbo 5.2) aproksimiramo matriko povratnih informacij R s produktom dveh, manjših matrik:

- *faktorska matrika uporabnikov* U, v kateri predstavlja vsaka vrstica enega od uporabnikov in
- *faktorska matrika vsebin* V, v kateri predstavlja vsaka vrstica eno od vsebin v sistemu za priporočanje.

$$R \approx X = U \cdot V' \tag{5.2}$$

Za reševanje problema hladnega zagona smo uporabili naslednje tehnike:

- Uporaba *povprečnih vrednosti faktorjev* v faktorski matriki (oznaka: "avgf") - Povprečne vrednosti faktorjev za nove uporabnike smo izračunali na podlagi vrednosti posameznih faktorjev v uporabniški faktorski matriki U, za nove vsebine pa na podlagi vrednosti v matriki V.

- *Faktorizacija razširjene matrike* povratnih informacij (oznaka: "mfex") - Pri tem postopku smo matriki povratnih informacij R dodali vrstico, ki predstavlja novega uporabnika. V to vrstico smo vpisali povprečne vrednosti interakcij s sistemom preko vseh uporabnikov, potem pa matriko faktorizirali. Problem hladnega zagona pri priporočanju novim uporabnikom smo reševali tako, da smo jim pripisali faktorje, ki pripadajo na novo ustvarjenemu uporabniku. Enak postopek smo uporabili tudi za izračun faktorjev za nove vsebine.

- *Napovedovanje faktorjev z reševanjem sistema linearnih enačb* (oznaka: "less") - Gantner [35] je predlagal uporabo semantičnih informacij za napovedovanje faktorjev v sistemih, ki temeljijo na matrični faktorizaciji. Za reševanje problema hladnega zagona sta predlagala reševanje sistema linearnih enačb, pri katerem za vsakega od faktorjev predvideva, da ga je možno aproksimirati z linearno kombinacijo vrednosti semantičnih atributov. V poskusih smo za gradnjo uporabniških profilov uporabili metodo AverageAction.

- *Napovedovanje faktorjev z uporabo metod strojnega učenja* (oznaki: "regt" in "nnet") - Pri evalvaciji Gantnerjevega pristopa [35] smo ugotovili, da bi se dalo preslikavo semantičnih informacij v faktorje izboljšati z uporabo metod strojnega učenja. Odločili smo se za uporabo regresijskih dreves, ki poleg hitrosti delovanja nudijo tudi dober vpogled v delovanje modela, in umetnih nevronskih mrež, s katerimi hkrati napovedujemo vrednosti več odvisnih spremenljivk. Pri napovedovanju faktorjev z regresijskimi drevesi smo za vsak faktor zgradili podatkovno množico s 130 vhodnimi atributi (uteži konceptov v profilih uporabnikov) in eno izhodno spremenljivko (faktor). Pri nevronskih mrežah smo zgradili le eno podatkovno množico s 130 vhodnimi atributi, odvisne spremenljivke pa so predstavljale posamezne faktorje. Za kalibracijo regresijskih dreves (velikost drevesa in ostalih parametrov) in umetnih nevronskih mrež (število

skritih nivojev in nevronov v njih ter stopnja regularizacije) smo uporabili 10-kratno prečno preverjanje.

Algoritem za matrično faktorizacijo

Za faktorizacijo matrike povratnih informacij R v uporabniško U in vsebinsko V faktorsko matriko uporabljamo uteženo regularizirano matrično faktorizacijo z izmenjujočimi najmanjšimi kvadrati WRMF (weighted regularized matrix factorization) [41]

Zaupanje v povratne informacije. Hu in sod. v članku [31] opisujejo algoritem za ma. fa., ki na podlagi implicitnih povratnih informacij uporabnika izračuna obe faktorski matriki, pri čemer pripiše vsaki pozitivni uporabnikovi povratni informaciji vrednost 1, negativni pa 0. Dodatno so v algoritem vpeljali pojem zaupanja v povratno informacijo C_{ui}, s pomočjo katerega lahko pri izračunu faktorskih matrik upoštevajo tudi negativne (oz. manjkajoče) povratne informacije. V enačbi 5.3 predstavlja R_{ui} vrednost celice v matriki povratnih informacij, α pa parameter za izračun zaupanja v to vrednost. Iz enačbe 5.3 sledi, da z večanjem vrednosti α povečujemo zaupanje v vrednosti v matriki povratnih informacij R. Vrednost parametra α je potrebno optimizirati na podatkovni množici. V članku [31] so zapisali, da so s pomočjo vrednosti $\alpha = 40$ dobili dobre rezultate, v naših poskusih pa smo najboljše rezultate dobili z vrednostjo $\alpha = 1.0$.

$$C_{ui} = 1 + \alpha R_{ui} \tag{5.3}$$

Regularizacija matrične faktorizacije. Algoritem, opisan v [31], omogoča tudi regularizacijo matrične faktorizacije, s katero se lahko izognemo morebitnemu prevelikemu prilagajanju učnim podatkom. V naših poskusih se je izkazalo, da pripelje uporaba regularizacije v postopku matrične faktorizacije do minimalnih razlik v faktorskih matrikah in posledično tudi v seznamih priporočil. Odločili smo se, da zato regularizacije ne bomo uporabljali.

Optimizacija vektorskih matrik. Metoda izmenjujočih najmanjših kvadratov ALS (angl. alternating least squares) je hitra in preprosta (brez dodatnih parametrov) metoda optimizacije, s katero izmenično optimiziramo vsako od vektorskih matrik. Pan in sod. [42] v članku opisujejo spremenjen optimizacijski postopek, poimenovan wALS (weighted alternating least squares), s katerim lahko damo med optimizacijo vektorskih

matrik pri izračunu Frobeniusove norme (t.j. kvadratne napake) aproksimizacijske matrike $L(X)$, katere izračun prikazuje enačba 5.4, večjo težo "polnim" celicam in manjšo težo "praznim" celicam matrike povratnih informacij.

$$L(X) = \sum_{ij} W_{ij}(R_{ij} - X_{ij})^2 \tag{5.4}$$

V enačbi 5.4 je R matrika povratnih informacij, X je aproksimacijska matrika, W pa matrika uteži. Za uteži v matriki W smo uporabili zaupanje (enačba 5.3) v vrednosti v matriki povratnih informacij $W = C$.

5.3.2 *Metrike za evalvacijo sistemov za priporočanje*

Sistemi za priporočanje se največkrat uporabljajo za dve nalogi:

1. gradnja seznama priporočil za uporabnika ali
2. napovedovanje uporabnikovih ocen za posamezne vsebine.

Od naloge je odvisen tudi način ocenjevanja uspešnosti sistemov za priporočanje. Za ocenjevanje delovanja sistemov za priporočanje se v raziskavah in na znanstvenih tekmovanjih običajno uporablja *koren povprečne kvadratne napake* (angl. root mean squared error) napovedi uporabnikovih ocen za posamezne vsebine.

V spletnem oglaševanju, kjer je največkrat edina eksplicitna povratna informacija s strani uporabnika klik na oglas, pa so bolj primerne metrike, ki ocenjujejo kvaliteto seznamov priporočil, pri čemer je najpomembnejše vprašanje, kako blizu začetka seznama so priporočila, ki jih uporabnik oceni pozitivno.

V poskusih na podatkih spletne oglaševalske mreže smo se omejili na oglede spletnih strani, pri katerih je prišlo do klika na oglas. Tako se v vsakem seznamu priporočil nahaja vedno le en uporaben predlog - to je oglas, na katerega je uporabnik kliknil.

Seznamov priporočil nismo ocenjevali na nivoju posameznih oglasov, temveč na podlagi tega, v katere oglasne skupine spadajo posamezni oglasi. Najpogostejša praksa v spletnem oglaševanju je namreč, da se vse oglase različnih velikosti in tipov, a z istim oglaševalskim sporočilom, za lažje obravnavanje združi v skupino oglasov. Razlog za ta način ocenjevanja tiči tudi v delovanju oglaševalske mreže, ki omogoča spletnim založnikom izbiranje tipov in velikosti oglasov, ki jih je dovoljeno prikazovati na njihovih spletnih straneh. Založniki se velikokrat odločijo, da bodo na primer na naslovnicah

dovolili le prikazovanje pasic (angl. banner) in slikovnih oglasov, tekstovni oglasi pa se lahko prikazujejo le ob vsebinskih člankih.

Za bolj natančno evalvacijo seznamov priporočil smo razvili metriki natančnost#N in priklic#N, ki za razliko od standardnih natančnost@N in priklic@N ne ocenjujeta dela seznama priporočil, temveč le posamezna priporočila v seznamu.

Povprečni rang

Povprečni rang (angl. average rank) izračunamo kot povprečno zaporedno številko uporabnih predlogov v seznamih priporočil (enačba 5.5). Ta metrika ocenjuje sisteme za priporočanje glede na zaporedno število "pravega" priporočila v seznamu priporočil. Je preprosta in lahko razumljiva, po drugi strani pa ne pove ničesar o distribuciji zaporednih števil pozitivnih priporočil v seznamih. Popolnoma enak rezultat bomo dobili na primer, če v prvem primeru izračunamo povprečen rang števil 1 in 100, ali pa 50 in 51.

$$rank_{avg} = \frac{\sum_{i=1}^{n} i}{n} \tag{5.5}$$

Natančnost@N in priklic@N

Standardni metriki klasifikacijska natančnost (enačba 5.6) in priklic (enačba 5.7) sta tu omejeni le na prvih N predlogov v seznamu. Pri manjših vrednostih N je tako večja verjetnost, da uporabnega predloga ne bomo našli med predlogi sistema za priporočanje.

Pri iskanju pozitivno ocenjenih priporočil v omejenih seznamih predlogov je pomembno, da se kliknjeni oglas nahaja čim višje v seznamu priporočil.

$$precision = \frac{items_{relevant} \cap items_{retrieved}}{items_{retrieved}} \tag{5.6}$$

$$recall = \frac{items_{relevant} \cap items_{retrieved}}{items_{relevant}} \tag{5.7}$$

Metrika DCG

Pri metriki DCG (angl. discounted cumulative gain) lahko vrednost relevance i-tega predloga rel_i v seznamu zavzame le vrednosti 0 (nerelevanten predlog) ali 1 (relevanten predlog).

$$rel_i = \begin{cases} 0 & ; rec_i \neq item_{relevant} \\ 1 & ; rec_i = item_{relevant} \end{cases} \tag{5.8}$$

Vrednost DCG (glej enačbo 5.9) se izračuna kot vsota relevanc vseh priporočil v seznamu, pri čemer so vrednosti relevanc zmanjšane v skladu z mestom i priporočila v seznamu. Ker upošteva metrika DCG pri izračunu vrstni red predlogov, je pri tej metriki pomembno, da se uporabni predlogi nahajajo čim bolj na čelu seznama priporočil.

$$DCG = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{log_2(i)}; rel_i \in 0,1 \tag{5.9}$$

Natančnost#N in priklic#N

Pri evalvaciji sistemov za priporočanje se veliko uporabljata metriki natančnost@N in priklic@N. Obe se osredotočata na prvih N priporočil sistema za priporočanje in ocenita kvaliteto tega dela seznama priporočil.

Za bolj natančno evalvacijo sistemov za priporočanje smo definirali metriki natančnost#N in priklic#N, ki izračunata natančnost in priklic N-tega priporočila v seznamu. Smiselnost uporabe natančnost#N je prikazana v razdelku 5.5, v katerem kombiniramo seznama priporočil dveh sistemov za priporočanje.

5.3.3 Evalvacija sistemov za priporočanje

V poskusih smo sisteme za priporočanje ocenjevali le na podlagi dogodkov, ko je uporabnik pozitivno ocenil enega izmed predlogov - to je klik na oglas. Teh dogodkov je bilo 7058. Ob vsakem kliku na oglas smo najprej zgradili množico aktivnih oglasov. Množice aktivnih oglasov so vsebovale med 443 in 567 oglasov.

Vsak sistem za priporočanje je aktivne oglase uredil v *seznam priporočil*, in sicer tako, da je na čelo seznama postavil oglas, ki naj bi bil najbolj primeren za danega uporabnika, na konec seznama pa najmanj primernega. V seznamu priporočil se je vedno nahajal tudi oglas, na katerega je uporabnik kliknil.

5.4 Poskusi: Primerjava osnovnih sistemov za priporočanje

V tem razdelku so predstavljeni rezultati evalvacij osnovnih sistemov za priporočanje (razdelek 5.3.1). V tabelah se pojavljajo naslednje okrajšave:

- *Rnd* - Naključna priporočila;
- *C/P* - Vsebinsko izbiranje, pri čemer zadnji člen *P* označuje profilirni algoritem, ki je bil uporabljen za pripravo učne množice za regresijska drevesa:
  - l - LastAction,
  - a - AverageAction,
  - afc - AverageActionFC,
  - d - Daoud2009,
  - g - Godoy2009,
  - s - Sieg2007;
- MF/*F*/*H* - Sistem za priporočanje na osnovi matrične faktorizacije (MF) s številom faktorjev *F*. Za reševanje problema hladnega zagona se uporablja tehnika *H*, ki lahko zavzame vrednosti:
  - avgf - novim uporabnikom se pripišejo povprečne vrednosti faktorjev iz faktorske matrike,
  - mfex - novim uporabnikom se pripišejo faktorji, dobljeni s faktorizacijo razširjene matrike povratnih informacij,
  - less - preslikava semantičnih informacij v faktorje z reševanjem sistema linearnih enačb,
  - regt - napovedovanje faktorjev z uporabo regresijskih dreves,
  - nnet - napovedovanje faktorjev z uporabo umetnih nevronskih mrež.

5.4.1 Optimizacija sistemov na osnovi matrične faktorizacije

Izbira primerne matrike povratnih informacij

Matrika povratnih informacij je zgrajena iz podatkov o interakcijah uporabnikov s sistemom za priporočanje. Ker tipičen uporabnik ne pride v stik z vsemi možnimi vsebinami, imamo opravka z *redko* matriko povratnih informacij. V naših poskusih smo opazili, da je "polnih" le okoli 5% celic matrike, kar sovpada z deležimi v sorodnih raziskavah.

Pri izbiri primerne velikosti matrike povratnih informacij smo se osredotočili na delovanje treh sistemov za priporočanje na osnovi matrične faktorizacije, ki vsak na svoj način rešuje problem hladnega zagona:

- MF/1/avgf - Novim uporabnikov se pripišejo povprečne vrednosti faktorjev iz faktorskih matrik,
- MF/1/regt - Za nove uporabnike določimo vrednosti faktorjev s pomočjo regresijskih dreves (učenje: atributi v profilih → faktor),
- MF/1/nnet - Za nove uporabnike določimo vrednosti faktorjev s pomočjo umetnih nevronskih mrež (učenje: atributi v profilih → faktorji).

Pri vseh treh sistemih smo omejili število faktorjev na $F = 1$, s čimer smo hoteli zagotoviti, da ne bi pri gradnji modela za napovedovanje faktorjev prihajalo do prevelikega prilagajanja učnim podatkom. Število uporabnikov, ki smo jih vključili v matriko povratnih informacij, smo povečevali postopoma. Najprej smo v matriko vključili uporabnike, katerih številka ID se začne z znakom "0", v naslednjem koraku smo dodali uporabnike, katerih številka ID se začne z znakom "1" itn.

Tabela 5.1

Povprečni rangi priporočil sistemov za priporočanje z uporabo različno velikih matrik povratnih informacij. V stolpcih so vrednosti za matrike s 1349, 2696, 4121 in 5508 uporabniki.

povp. rang	1349	2696	4121	5508
MF/1/avgf	23.67	20.19	20.08	19.97
MF/1/regt	23.42	20.76	20.51	19.87
MF/1/nnet	19.34	18.93	18.67	18.47

V tabelah 5.1 - 5.3 vidimo, da lahko kvaliteto priporočil z uporabo matrike iz interakcij 1349 uporabnikov bistveno izboljšamo, če povečamo število zajetih uporabnikov na 2696. Razlike med matrikami velikosti 2696, 4121 in 5508 uporabnikov so majhne, kvaliteta predlogov v tabelah 5.2 in 5.3 pa ne narašča vedno s povečevanjem števila uporabnikov. Iz tega lahko sklepamo, da je za dobro priporočanje vsebin v tem primeru dovolj velika že matrika 2696 uporabnikov.

Tabela 5.2

Prikaz vrednosti natančnost@10 za sisteme za priporočanje z uporabo različno velikih matrik povratnih informacij. V stolpcih so vrednosti za matrike s 1349, 2696, 4121 in 5508 uporabnikov.

natančnost@10	1349	2696	4121	5508
MF/1/avgf	0.4906	0.5831	0.5794	0.5797
MF/1/regt	0.4896	0.5703	0.5747	0.5818
MF/1/nnet	0.6055	0.6114	0.6138	0.6090

Tabela 5.3

Prikaz vrednosti DCG@ 10 za sisteme za priporočanje z uporabo različno velikih matrik povratnih informacij. V stolpcih so vrednosti za matrike s 1349, 2696, 4121 in 5508 uporabnikov.

DCG@10	1349	2696	4121	5508
MF/1/avgf	903.8	1491.1	1481.1	1479.9
MF/1/regt	891.5	1406.0	1416.0	1456.8
MF/1/nnet	1201.3	1287.4	1280.5	1238.2

Izbira števila faktorjev

Algoritmu matrične faktorizacije je potrebno vnaprej določiti število faktorjev v faktorskih matrikah. V tabeli 5.4 vidimo, da je en faktor premalo za gradnjo dobrih seznamov priporočil, že uporaba treh faktorjev pa močno izboljša rezultate. Najboljše rezultate smo dobili z uporabo 20 faktorjev $F = 20$.

5.4.2 Rezultati

V tabeli 5.5 vidimo, da se glede na povprečni rang najbolje odreže sistem za priporočanje z vsebinskim izbiranjem, ki gradi profil spletnega uporabnika z metodo Sieg2007 [8]. Med sistemi na osnovi matrične faktorizacije v tem pogledu ni velikih razlik, v negativnem smislu odstopa le pristop avtorja Gantner [35].

Sistema “Rand” in “MF/20/nnet” sta edina nedeterministična. Oba smo evalvirali 10-krat, v tabelah pa so predstavljeni povprečni rezultati.

Metrika natančnost@N ocenjuje kvaliteto omejenih seznamov priporočil. Rezultati v tabeli 5.6 kažejo, da so sistemi na osnovi matrične faktorizacije veliko boljši pri predlaganju kratkih seznamov priporočil. Vrednosti natančnost@1 ≈ 0.5 pomenijo, da je bilo v povprečju vsakič drugič pozitivno ocenjeno prvo priporočilo v seznamu, kar

Tabela 5.4

Prikaz vrednosti metrik povprečni rang, natančnost@10 in DCG@10 za sisteme za priporočanje pri uporabi različnega števila faktorjev.

	povp. rang	natančnost@10	DCG@10
MF/1/regt	20.76	0.5703	1406.0
MF/3/regt	10.34	0.7419	2100.6
MF/5/regt	10.62	0.7385	2168.2
MF/10/regt	10.05	0.7549	2117.2
MF/20/regt	*9.63*	*0.7649*	*2313.9*
MF/30/regt	9.72	0.7602	2291.5
MF/1/nnet	18.93	0.6114	1287.4
MF/3/nnet	28.27	0.4585	1523.7
MF/5/nnet	17.21	0.6439	1768.3
MF/10/nnet	*9.42*	0.7703	2248.0
MF/20/nnet	9.58	*0.7785*	*3486.5*
MF/30/nnet	9.69	0.7692	3327.4

Tabela 5.5

Primerjava rezultatov osnovnih sistemov za priporočanje z metrikami povprečni rang, natančnost@10 in DCG@10.

	povp. rang	natančnost@10	DCG@10
Rnd	39.79	0.0793	230.0
C/l	12.29	0.5407	1586.2
C/a	11.02	0.6349	2072.5
C/afc	9.60	0.7277	2741.8
C/d	9.88	0.6814	2776.5
C/g	11.55	0.5941	1853.6
C/s	8.74	0.7663	3292.4
MF/20/avgf	9.64	0.7615	2281.9
MF/20/less	19.10	0.1649	346.8
MF/20/regt	9.63	0.7604	2305.8
MF/20/nnet	9.57	0.7785	3486.4

se sliši neverjetno. Zelo visoke vrednosti v tabeli 5.6 so posledica dejstva, da so spletni uporabniki v učnem obdobju zelo dobro sprejeli in veliko klikali na manjše število oglasov, to obnašanje pa se je nadaljevalo tudi v testnem obdobju našega poskusa.

Pri manj omejenih seznamih se razlika med vsebinskim izbiranjem in izbiranjem na podlagi sodelovanja zmanjša, pri vrednosti $N = 20$ pa že vidimo, da se bolj obnese vsebinsko izibiranje.

Tabela 5.6

Primerjava rezultatov osnovnih sistemov za priporočanje z metriko natančnost@N.

	nat@1	nat@3	nat@5	nat@10	nat@20
Rnd	0.0068	0.0212	0.0379	0.0793	0.1786
C/l	0.0029	0.0863	0.1879	0.5407	0.8512
C/a	0.0032	0.1355	0.3073	0.6349	0.8544
C/afc	0.0010	0.2430	0.4583	0.7277	0.8628
C/d	0.0015	0.2768	0.5008	0.6814	0.8549
C/g	0.0042	0.1138	0.2559	0.5941	0.8524
C/s	0.0012	0.3524	0.6071	0.7663	0.8632
MF/20/avgf	0.3240	0.6354	0.7429	0.7615	0.7800
MF/20/less	0.0003	0.0027	0.0101	0.1649	0.7166
MF/20/regt	0.3242	0.6477	0.7434	0.7604	0.7783
MF/20/nnet	0.1177	0.6098	0.7491	0.7752	0.7860

Pri metriki DCG@N je še bolj pomembno, da se pozitivno ocenjena priporočila nahajajo čimbolj na začetku seznamov priporočil. V tabeli 5.7 zopet vidimo, da se pri kratkih seznamih priporočil bolj odrežejo sistemi na osnovi matrične faktorizacije, pri daljši seznamih pa je razlika manjša. Med sistemi z vsebinskim izbiranjem smo najboljše rezultate dobili s profilirnim algoritmom Sieg, ki se marsikje odreže bolje od večine sistemov na osnovi matrične faktorizacije.

5.5 *Poskusi: Kombiniranje priporočil različnih sistemov za priporočanje*

Iz rezultatov v razdelku 5.4.2 vidimo, da so v splošnem sistemi z izbiranjem na podlagi sodelovanja boljši pri predlaganju zelo kratkih seznamov priporočil, sistemi z vsebinskim izbiranjem pa pri dolgih seznamih. To pomeni, da je na začetkih seznamov

Tabela 5.7

Primerjava rezultatov osnovnih sistemov za priporočanje z metriko DCG@N.

	DCG@1	DCG@3	DCG@5	DCG@10	DCG@20
Rnd	43	115	161	247	398
C/l	208	637	1015	1684	2091
C/a	270	1073	1572	2139	2434
C/afc	572	1731	2410	2803	2966
C/d	663	1993	2554	2827	3064
C/g	239	872	1322	1932	2277
C/s	847	2494	3148	3313	3450
MF/20/avgf	1056	2240	2252	2291	2312
MF/20/less	2	22	37	347	1221
MF/20/regt	*1137*	2256	2280	2314	2336
MF/20/nnet	696	*2910*	*3424*	*3480*	*3498*

priporočil sistemov z izbiranjem na podlagi sodelovanja nekaj zelo dobrih priporočil, ostali pa so slabši. Pri vsebinskem izbiranju so začetna priporočila relativno slabša, je pa padec kvalitete priporočil z zaporedno številko priporočila dosti manjši.

Eden od vzrokov za slabša začetna priporočila sistemov z vsebinskim izbiranjem je kombinacija kontekstualnega oglaševanja, ki je bilo uporabljeno v testnem obdobju, in umeščanje oglasov na spletne strani. O tem je več napisanega v razdelku B.

Ker so priporočila na začetkih seznamov priporočil sistemov "MF/20/avgf" in "MF/20/regt" bolj kvalitetna kot pri "MF/20/less" in "MF/20/nnet", smo slednja izpustili iz nadaljnjih poskusov. Zaradi večje preglednosti prikazujemo v tem razdelku le rezultate sistemov za priporočanje, ki uporabljajo profilirni algoritem AverageAction.

Tabela 5.8 prikazuje kvaliteto priporočil glede na njihovo zaporedno številko v seznamu. Vidimo, da so pri priporočilih z zaporednimi številkami $1 \leq i \leq 4$ boljši sistemi z izbiranjem na podlagi sodelovanja, pri nadaljnjih priporočilih pa sistemi z vsebinskim izbiranjem.

V skladu s tabelo 5.8 smo na podlagi parov sistemov za priporočanje zgradili kombinirane sezname priporočil. V vsakem paru sistemov sta po en sistem z izbiranjem na podlagi sodelovanja in en sistem z vsebinskim izbiranjem. Kombiniran seznam priporočil smo zgradili tako, da smo prvih nekaj priporočil vzeli iz sistema z izbiranjem na

Tabela 5.8

Primerjava rezultatov osnovnih sistemov za priporočanje z metriko natančnost#N.

	nat#1	nat#2	nat#3	nat#4	nat#5
C/a	0.0032	0.0457	0.0866	0.0781	0.0937
MF/20/avgf	0.3240	0.1786	0.1328	0.1071	0.0005
MF/20/regt	0.3242	0.1923	0.1312	0.0918	0.0039

podlagi sodelovanja, seznam pa smo potem dopolnili z priporočili sistema z vsebinskim izbiranjem. Oba osnovna seznama priporočil vsebujeta enake elemente (razlikujeta se le v vrstnem redu elementov), zato je tudi kombinirani seznam enako dolg.

V tabelah 5.9 in 5.10 so prikazani rezultati obeh osnovnih in kombiniranega sistema za priporočanje. V obeh primerih smo v skladu s tabelo 5.8 iz seznama priporočil sistema z izbiranjem na podlagi sodelovanja vzeli prva 4 priporočila, za katere lahko zaradi visokih izkazanih vrednosti natančnost#1 predvidevamo, da so zelo kvalitetna. Seznam smo dopolnili s priporočili iz sistema z vsebinskim izbiranjem. Rezultati kažejo, da so kombinirani seznami priporočil občutno boljši od priporočil osnovnih sistemov za priporočanje.

Tabela 5.9

Primerjava rezultatov kombiniranih sistemov za priporočanje. V kombiniranem seznamu priporočil so prva 4 priporočila iz sistema "MF/20/avgf", ostala pa iz "C/a".

	povp. rang	natančnost@10	DCG@10
C/a	11.02	0.6349	2139
combined	7.08	0.7871	2343
MF/20/avgf	9.64	0.7615	2291

5.6 Ugotovitve

Pri evalvaciji sistemov za priporočanje nam najprej pade v oči nenavadno slaba kvaliteta priporočil sistemov z vsebinskim izbiranjem. Eden od vzrokov za to je kombinacija kontekstualnega oglaševanja in umeščanja oglasov v spletne strani, o čemer je več napisanega v razdelku B. Glede na to, da je vsebovalo v naših poskusih okoli 80% spletnih strani le en oglasni blok, predvidevamo, da botruje slabim rezultatom sistemov z vse-

Tabela 5.10

Primerjava rezultatov kombiniranih sistemov za priporočanje. V kombiniranem seznamu priporočil so prva 4 priporočila iz sistema "MF/20/regt", ostala pa iz "C/a".

	povp. rang	natančnost@10	DCG@10
C/a	11.02	0.6349	2139
combined	7.05	0.7874	2366
MF/20/regt	9.63	0.7604	2314

binskim izbiranjem več vzrokov. Iskanje možnih vzrokov bo predmet nadaljnjega dela.

V poglavju 4 smo pokazali, da je metoda AverageActionFC boljša od obstoječih metod za gradnjo ontoloških profilov, zato smo pričakovali, da bomo lahko to potrdili tudi pri uporabi uporabniških profilov v sistemih za priporočanje. Neugodna kombinacija kontekstualnega oglaševanja in umeščanja oglasov v spletne strani bi lahko razložila dejstvo, da je bila najslabša kvaliteta "prvih" priporočil izmerjena ravno ob uporabi naše profilirne metode AverageActionFC, za katero smo v prejšnjem poglavju pokazali, da lahko z njo zgradimo najbolj kvalitetne uporabniške profile.

Reševanje problema hladnega zagona pri sistemih za priporočanje, ki uporabljajo matrično faktorizacijo, se navadno rešuje z uporabo dodatnih virov informacij (uporabnikova lokacija, profil itd.). Preslikava semantičnih informacij v faktorje z uporabo sistema linearnih enačb, kot je predlagal Gantner [35], se na naši podatkovni množici ni obneslo. Verjeten vzrok za to je preobčutljivost tega pristopa na šum v podatkih. Preslikava z uporabo metod strojnega učenja (regresijska drevesa in umetne nevronske mreže) se je sicer izkazala za koristno, vendar je doprinos zelo majhen.

Temeljita analiza kvalitete priporočil je pokazala očiten trend, in sicer da dajejo sistemi z izbiranjem na osnovi sodelovanja nekaj zelo dobrih priporočil, čemur sledi velik padec kvalitete. Priporočila vsebinskega izbiranja po drugi strani ne dosegajo kvalitete najboljših priporočil sistemov z izbiranjem na osnovi sodelovanja, prav tako pa ni prisoten tako očiten padec kvalitete priporočil. Na podlagi teh informacij smo zgradili kombinirane sezname priporočil, pri katerih je opaziti velik preskok v kvaliteti.

Zaključek

6

V disertaciji smo se ukvarjali z gradnjo, evalvacijo in uporabo uporabniških profilov v sistemih za priporočanje. V poskusih smo uporabljali podatke spletne oglaševalske mreže in spletne učilnice, osnovane na sistemu Moodle. Analiza rezultatov je pokazala, da so za gradnjo kvalitetnih profilov spletnih uporabnikov potrebni dolgi učni klikotoki. Ugotovili smo še, da se s staranjem profilov njihova kvaliteta znižuje. Profile spletnih uporabnikov, ki smo jih zgradili na podlagi njihovih klikotokov, smo izboljšali z novim algoritmom AverageActionFC, ki uporablja tehniki *časovnega pozabljanja* in *popravljanja profilov s prototipi*. Primerjava z obstoječimi metodami je pokazala, da je kvaliteta profilov, zgrajenih z algoritmom AverageActionFC značilno boljša od profilov, zgrajenih z obstoječimi metodami.

Nizka kvaliteta uporabnikovega profila je lahko posledica starosti profila ali pomanjkanja uporabnih informacij o uporabnikovih preteklih aktivnosti. Takim profilom smo uspeli občutno zvišati kvaliteto z uporabo popravljanja s prototipi. Pri evalvaciji profilov na obeh uporabljenih podatkovnih množicah se je izkazalo, da je optimalna vrednost uteži za popravljanje profilov relativno visoka ($0.7 \leq w_{correct} \leq 0.9$). To pomeni, da je imelo na uporabnikov profil prevladujoč vpliv domensko znanje (v obliki prototipnih profilov) in ne njegov lastni klikotok.

Profili spletnih uporabnikov so največkrat uporabljeni v sistemih za priporočanje. Podatke spletne oglaševalske mreže smo uporabili za evalvacijo sistemov za priporočanje z vsebinskim izbiranjem in izbiranjem na podlagi sodelovanja. Profile spletnih uporabnikov smo gradili na podlagi njihovih obiskov spletnih strani, za priporočila pa smo uporabili oglase, ki so se prikazovali uporabnikom. Rezultati kažejo, da so bili oglasi, predlagani z vsebinskim izbiranjem, zelo redko kliknjeni. Ugotovili smo, da je eden od vzrokov za take rezultate strategija izbiranja oglasov za prikazovanje, ki se je uporabljala v testnem obdobju. Pri kontekstualnem oglaševanju se namreč za najbolj primerne oglase izberejo tisti, ki se najbolje ujemajo z vsebino obiskane spletne strani. Analiza klikov na oglase in njihovih lokacij na spletnih straneh je pokazala, da ima lokacija oglasa velik vpliv na njegovo uspešnost. Kljub temu, da so se oglasi na vrhu in na levi strani spletne strani najbolj ujemali z njeno vsebino, so bili ravno ti najmanjkrat kliknjeni.

Priporočila sistemov z izbiranjem na podlagi sodelovanja so se izkazala za boljša. Pri izbiranju na podlagi sodelovanja pa se v spletnem okolju ne moremo izogniti problemu hladnega zagona. Za reševanje tega problema smo v sistemih, ki temeljijo na matrični faktorizaciji, uporabili regresijska drevesa in umetne nevronske mreže za *napovedovanje*

faktorjev novih uporabnikov in vsebin na podlagi razpoložljivih semantičnih informacij. Rezultati so pokazali, da lahko s tem pristopom omilimo negativne posledice hladnega zagona.

Podrobnejša analiza kvalitet priporočil sistemov za priporočanje je pokazala, da dajejo sistemi z izbiranjem na podlagi sodelovanja le nekaj zelo dobrih priporočil, priporočila sistemov z vsebinskim izbiranjem pa izkazujejo nekoliko nižjo, a bolj stabilno raven kvalitete. S *pametnim kombiniranjem seznamov priporočil* smo uspeli zgraditi občutno boljše sezname priporočil, kar lahko pomembno prispeva k znanstveni in ekonomski uspešnosti sistemov za priporočanje.

6.1 Nadaljnje delo

Velik problem pri gradnji profilov predstavljata hramba in obdelava velike količine podatkov. Večjo hitrost in prostorsko učinkovitost bi lahko dosegli z uporabo metod, ki omogočajo inkrementalno posodabljanje uporabniških profilov. V algoritmu AverageActionFC nameravamo preizkusiti še t.i. dinamične prototipne profile, ki jih bomo definirali s pomočjo inkrementalnega razvrščanja uporabniških profilov. To nam bo omogočilo, da se množica prototipnih profilov sproti prilagaja obnašanju in navadam celotne populacije spletnih uporabnikov. Z uporabo drugih algoritmov razvrščanja bomo lahko tudi poenostavili postopek definiranja množice prototipov, saj za razliko od zdaj uporabljenega k-means večina algoritmov razvrščanja ne zahteva vnaprejšnje določitve števila gruč k.

Pri gradnji ontoloških profilov smo uporabljali le ročno zgrajene ontologije, zato bi bilo smiselno v poskuse vključiti še avtomatsko ali polavtomatsko zgrajene ontologije, npr. z metodo OntoGen.

V poskusih na podatkih spletne oglaševalske mreže se je pokazalo, da lahko kvaliteto dolgoročnih napovedi izboljšamo tudi s popravljanjem profilov s časovnimi statistikami, ki odsevajo popularnost tematik ob različnih dnevih in urah. Ta pristop se je sicer izkazal za slabšega od algoritma AverageActionFC, ki uporablja popravljanje s prototipnimi profili, verjamemo pa, da bi lahko kvaliteto profilov dodatno izboljšali s kombinacijo obeh tehnik.

Problem hladnega zagona smo v poskusih na podatkih oglaševalske mreže reševali z napovedovanjem na podlagi semantičnih informacij o novih uporabnikih in vsebinah. Verjamemo, da je mogoče hladni zagon rešiti še bolj učinkovito, npr. z uporabo uporabnikove trenutne lokacije in drugih informacij, ki so lahko na razpolago v spletnem

okolju. Sistemi z izbiranjem na podlagi sodelovanja so med drugim občutljivi tudi na usmerjene napade uporabnikov na delovanje sistema. V nadaljevanju se bomo zato posvetili tudi odpornosti sistemov na t.i. shilling napade, gray sheep napade ipd.

Glede na izsledke poskusov z metodo, ki popravlja uporabniške profile s časovnimi statistikami (razdelek 3.2.1), se zdi smiselno v proces priporočanja uporabnikom vključiti tudi časovno komponento. S tem bi lahko ob določenih urah okrepili prikazovanje oglasov, ki so vezani na takrat bolj popularne tematike, kar bi lahko povečalo uspešnost oglaševanja. S takim pristopom ali pa z detekcijo semantičnih preskokov v obiskanih spletnih straneh bi lahko tudi bolje modelirali interese uporabnikov, ki uporabljajo družinske računalnike.

S podatki oglaševalske mreže smo bili v poskusih omejeni le na pozitivne povratne informacije uporabnikov o vsebinah – kliki na oglase. Verjamemo, da bi lahko imela vključitev negativnih povratnih informacij pozitiven vpliv na kvaliteto priporočil, zato bomo vložili dodaten napor v zbiranje tovrstnih informacij.

Pri profiliranju uporabnikov in uporabi njihovih osebnih podatkov v sistemih za priporočanje je potrebno spoštovati zasebnost in pravico uporabnikov do izločitve njihovih podatkov iz analiz. Tej zelo aktualni problematiki se bomo aktivneje posvetili v prihodnosti.

Še posebej je s stališča zasebnosti pomemben način sledenja spletnim uporabnikom. Raziskave so pokazale, da je mogoče na podlagi podatkov o uporabnikovem računalniku (operacijski sistem, brskalnik, seznam vtičnikov in nameščenih pisav, ipd.) zelo natančno razločevati med uporabniki. Zanimivo bi bilo preveriti, kako dobro bi se dalo priporočati spletnim uporabnikom le na podlagi takih podatkov.

Dodatek: Zasebnost na spletu

A

A.1 Kratka zgodovina oglaševanja

Prvi oglasi so se med ljudmi širili verbalno in slikovno. V antiki so trgovci na ves glas oglaševali svoje blago, glasniki pa prihod ladij in trgovskih karavan. V Starem Egiptu so se z razvojem pisave in papirusa pojavili prvi plakati s prodajnimi sporočili. V ostankih Pompejev so na nekaterih stenah hiš premožnih meščanov našli sledove komercialnih napisov in celo političnih oglasov.

Zaradi visoke stopnje nepismenosti med ljudstvom so obrtniki še v srednjem veku svoje delavnice namesto z napisi označevali s simboli (čevljarji s čevljem, kovači z nakovalom ali podkvijo, mlinarji z vrečo moke itd.). Ostali trgovci, ki so svoje blago prodajali z vozov, pa so svojo lokacijo morebitnim kupcem sporočali prek mestnih klicarjev.

Oglaševanje je vedno zvesto sledilo novim tehnologijam. Razvoj tiska je omogočil oglaševanje v časopisih in na plakatih. Konec 19. stoletja je Thomas J. Barratt, znan tudi kot oče modernega oglaševanja, kot prvi začel z oglaševanjem izdelka s kombinacijo fraz, sloganov, slik in pričevanj znanih oseb ter znanstvenikov.

Z razmahom industrializacije in poplave izdelkov na tržišču se je razvilo tudi masovno oglaševanja, katerega namen je povečanje povpraševanja in potrošnje. V tem času se je uveljavila teorija, da lahko oglaševalci z izkoriščanjem človeških nagonov (npr. potreba po pripadanju, ljubezni, spoštovanju, ugledu, prepoznavnosti v družbi, narcisizem) povečajo željo po nakupu določenih izdelkov.

Izum televizije je sprožil umik oglaševalskega denarja z zunanjih medijev (plakati ipd.) na televizijske kanale. Večanje števila in raznolikost televizijskih kanalov pa je kasneje razdelila ciljno publiko in zmanjšala učinkovitost oglaševanja na televiziji, ter s tem ponovno oživela oglaševanje na zunanjih in tiskanih medijih.

Komercializacija in popularizacija interneta je v zadnjem desetletju prejšnjega stoletja omogočila oglaševalcem nove, kreativne prijeme, fleksibilnost, hitrost, dostopnost ciljne publike po vsem svetu in učinkovito merjenje uspešnosti oglaševanja. Prvi oglas, posredovan preko interneta, je bil v obliki nezaželene elektronske pošte poslan že maja 1978. Leta 1993 se je pojavila prva oglaševalska pasica, sedaj pa smo spletni uporabniki stalno izpostavljeni bolj ali manj vsiljivim oglasom pri večini aktivnosti, povezanih z internetom: med prebiranjem pošte, novic, med uporabo spletnih iskalnikov, na socialnih omrežjih itd. V podjetju Google so prvi začeli s t.i. kontekstualnim oglaševanjem, v okviru katerega se uporabnikom prikazujejo nevsiljivi oglasi, ki so vsebinsko

povezani z obiskano spletno stranjo.

V zadnjem času se je pojavilo t.i. gverilsko oglaševanje, pri katerem poskušajo oglaševalci s čimbolj poceni, z zabavnimi, šokantnimi ali celo ilegalnimi pristopi ujeti pozornost ciljne publike. Ta tip oglaševanja je zaradi cenovne dostopnosti in nenavadnosti postal popularen predvsem pri manjših podjetjih, ki si skušajo na ta način dvigniti prepoznavnost.

A.2 Pravni in etični vidiki oglaševanja

V Evropski uniji področje oglaševanja ni podrobno urejeno. Nekatere omejitve in navodila, katerih namen je postopna izenačitev dovoljenih načinov oglaševanja znotraj EU, so državam članicam podane v obliki direktiv in uredb, sicer pa je vsaka država odgovorna za sprejem primerne zakonodaje.

V Sloveniji je oglaševanje regulirano s številnimi zakoni (Zakon o varstvu konkurence, Zakon o medijih, Zakon o varstvu potrošnikov itd.) in predpisi. Oglaševalci, ki se borijo za pozornost publike, pogosto hodijo po tanki črti med zakonitim in nezakonitim, še pogosteje pa prestopajo mejo dobrega okusa. Ta meja je zelo subjektivna, v Slovenski oglaševalski zbornici pa so jo poskušali definirati v Slovenskem oglaševalskem kodeksu [43] (SOK). To orodje samoregulacije definira pravila oglaševanja v Sloveniji, nima pa nobene zakonske moči.

Še posebej v zadnjih letih se v medijih pojavljajo oglasi, ki vsaj na prvi pogled prestopajo meje dovoljenega.

Šokantno oglaševanje

Šokantno oglaševanje poskuša pritegniti pozornost občinstva s prikazovanjem šokantnih prizorov in računa na čustven odziv gledalcev. Taki oglasi lahko kršijo zakonodajo zaradi nedostojnosti, po mnenju nekaterih pa je sporna tudi neočitna povezava med oglasom in oglaševanim izdelkom oz. storitvijo.

Primerjalno oglaševanje

O *primerjalnem oglaševanju* govorimo, če v oglasu izrecno ali z nakazovanjem med seboj primerjamo dva ali več izdelkov oz. storitev, pri tem pa nedvoumno identificiramo konkurenta. V Sloveniji je tak način oglaševanja dovoljen le pod določenimi pogoji (objektivna primerjava cen in lastnosti, ne ustvarja zmede, ne diskreditira ali očrni konkurenta, obravnava izdelke z enakim poreklom itd.). V našem medijskem prostoru

takega oglaševanja niti ni zaznati, še najbližje pridejo oglasi, kjer se izdelek primerja z generičnim konkurentom brez kakršnihkoli oznak.

Povsem drugače je v ZDA in Veliki Britaniji, kjer se velika podjetja rada zapletajo v neposredne primerjave in celo žaljenja. Lep primer je oglas za Adidas[1], ki prikazuje tekača, ki teče po težavnem terenu. Po dobrih desetih sekundah nam v oglasu razložijo, da ima tekač obute superge znamke Nike, kamerman, ki teče poleg tekača in nosi še težko kamero, pa nosi superge Adidas.

Suženjsko posnemanje

Suženjsko posnemanje namenoma zakriva razlike med oglaševanim izdelkom oz. storitvijo, in konkurenčnim, ki je ponavadi bolj prepoznaven. Pri suženjskem posnemanju prepoznavnega izdelka lahko pride do kršenja avtorskih pravic, kar lahko vodi v pravni spor.

Leta 1997 sta se v sodni spor zapletli podjetji Pivovarna Laško d.d. in Tara d.o.o. Slednje je na trgu ponujalo pivo Kozorog, katerega embalaža je bila na moč podobna Zlatorogu iz Pivovarne Laško. Sodišče druge stopnje je podjetju Tara d.o.o. prepovedalo uporabo blagovnega znaka sestavljenega iz besed: "Kozorog pivo" in s silhueto kozoroga v ovalu s stiliziranimi gorami v ozadju na steklenicah piva [44].

Stereotipiziranje

Prikazovanje stereotipov je verjetno eden najlažjih in najbolj razširjenih orodij za izdelavo oglasov. Taki oglasi izkoriščajo znane stereotipe za hitrejši prenos sporočila do ciljne publike, največkrat skozi humor. Posamezniki so v oglasih prikazani z le eno karakteristiko, to pa je lahko za marsikoga žaljivo.

V oglasih so ženske še vedno pogosto prikazane le kot gospodinje ali mame, moški kot karieristi ali avanturisti, starejši pa imajo probleme z zdravjem in moderno tehnologijo. Poleg spolne pripadnosti in starosti so viri stereotipov lahko še narodnost, rasa, veroizpovedi itd.

Zaščita otrok in mladoletnikov

Zakon o medijih se med drugim ukvarja tudi z oglasi, ki ciljajo na otroke in mladoletnike. Ti so zaradi neizkušenosti lahka tarča oglaševalcev, na udaru pa so seveda denarnice njihovih staršev. Oglasi, ki so namenjeni tej populaciji, ne smejo vsebovati

[1] Nike Vs Adidas - YouTube https://www.youtube.com/watch?v=8TnG7jyfoWI

neprimernih prizorov, izkoriščati zaupanja otrok v starše in učitelje, nagovarjati otroke k nakupu ali prepričevanju staršev v nakup ali jih neupravičeno prikazovati v nevarnih situacijah.

Dodatna pravila v Slovenskem oglaševalskem kodeksu [43] pa še prepovedujejo:

- zmanjševanje pomena cene z besedami "samo", "le",
- spodbujanje slabih prehrambenih navad,
- prikazovanje otrok, ki se nagibajo čez okna ali ograje mostov ipd.

A.3 Zasebnost spletnih uporabnikov

Pod pojmom *zasebnosti spletnih uporabnikov* se največkrat razume pravica uporabnika do nadzora nad zbiranjem, obdelovanjem in prikazovanjem podatkov o njem.

Varovanje zasebnosti spletnih uporabnikov je v zadnjih letih postala zelo aktualna tema. Spletni uporabniki se vedno bolj zavedajo svojih pravic in pomembnosti varovanja osebnih podatkov [45], ki se lahko preko spleta hitro in nekontrolirano razširjajo po svetu.

Največ osebnih podatkov o spletnih uporabnikih imajo brez dvoma na voljo socialna omrežja, ki jim uporabniki pogosto zaupajo svoja imena, starost, spol, najljubše glasbene skupine, filme, blagovne znamke, pa tudi bolj kočljive informacije, kot je na primer politična pripadnost, rasa ali veroizpoved. Predvsem mladi, ki uporabljajo socialna omrežja za vsakodnevno komunikacijo, dostikrat niso pozorni na vsebino njihovih objav. Na socialnih omrežjih objavljajo tudi kočljive slike z zabav, brisanje slik pa pri večini spletnih aplikacij implementirano tako, da je slika sicer umaknjena s spletnih strani, kopija slike pa ostane v podatkovni bazi podjetja. V zadnjem času se veliko dela tudi na ozaveščanju mladih, čemur je namenjena spletna stran Safe-si[2]. Največje socialno omrežje Facebook je v zadnjih letih razvilo zelo napreden algoritem za prepoznavanje obrazov, ki deluje celo boljše od algoritma ameriške agencije FBI[3].

Ogromno informacij o uporabnikih imajo na voljo tudi ostala velika spletna podjetja, kot so Google, Yahoo!, Apple in Microsoft. Uporabnikom ta podjetja ponujajo brezplačne ali poceni storitve, kot so poštni predal, aplikacije za takojšnje sporočanje (angl. instant messaging) in prostor za uporabnikove datoteke na njihovih strežnikih.

[2] Safe-si, varna raba interneta http://safe.si/

[3] http://www.usnews.com/news/articles/2014/07/08/fbi-may-seek-facebook-data-for-facial-recognition

Bolj omejen dostop do uporabniških informacij imajo običajne spletne strani, ki lahko spremljajo uporabnikove aktivnosti v le znotraj spletnega mesta, uporabniki pa se ponavadi predstavljajo s psevdonimi. Podobno velja tudi za oglaševalske mreže, ki pa lahko uporabnika z uporabo spletnih piškotkov ali podobnih metod sledijo uporabnikovim aktivnostim na vseh spletnih straneh, ki so dale mreži v najem prostor za prikazovanje oglasov. Take zbirke informacij so zlata jama tako za zasebna podjetja kot tudi za vladne organizacije, predvsem organe pregona. Zasebna podjetja največkrat uporabijo informacije o uporabnikih za bolj učinkovito oglaševanje, nekatera pa jih prodajajo naprej drugim podjetjem ali vladnim organizacijam. Organi pregona in druge vladne organizacije ponavadi upravičujejo zbiranje informacij o državljanih z bojem proti kriminalu, kljub temu pa so pavšalni posegi v zasebnost državljanov brez utemeljenega suma in naloga sodišča najmanj nemoralni.

A.3.1 Ustavnopravna in zakonska zaščita spletnih uporabnikov

Ustava Republike Slovenije

Ustava je najvišji pravni akt Republike Slovenije, kar pomeni, da morajo biti vsi zakoni in drugi pravni akti v skladu z Ustavo. Za varovanje zasebnosti spletnih uporabnikov sta najpomembnejša 37. in 38. člen Ustave RS, ki varujeta tajnost komunikacij in osebnih podatkov [46].

Kljub temu, da morajo biti vsi zakoni v skladu z Ustavo, pa je teoretično možno, da vlada z zakonom ratificira mednarodno pogodbo, ki je v delno ali v celoti v neskladju z Ustavo [47]. V tem primeru je lahko država postavljena v neprijeten položaj, ko je prisiljena izbirati med kršenjem lastne Ustave ali pa mednarodne pogodbe, s čimer lahko stori mednarodni delikt. V zadnjih letih so veliko prahu dvignile informacije o tajnih pogajanjih o mednarodnih trgovinskih sporazumih ACTA[4] in TTIP[5], ki bi v primeru ratifikacije občutno znižala nivo zaščite zasebnosti spletnih uporabnikov.

[4] Anti-Counterfeiting Trade Agreement

[5] Transatlantic Trade and Investment Partnership

37. člen Ustave Republike Slovenije (varstvo tajnosti pisem in drugih občil)

Zagotovljena je tajnost pisem in drugih občil.

Samo zakon lahko predpiše, da se na podlagi odločbe sodišča za določen čas ne upošteva varstvo tajnosti pisem in drugih občil in nedotakljivost človekove zasebnosti, če je to nujno za uvedbo ali potek kazenskega postopka ali za varnost države.

38. člen Ustave Republike Slovenije (varstvo osebnih podatkov)

Zagotovljeno je varstvo osebnih podatkov. Prepovedana je uporaba osebnih podatkov v nasprotju z namenom njihovega zbiranja.

Zbiranje, obdelovanje, namen uporabe, nadzor in varstvo tajnosti osebnih podatkov določa zakon.

Vsakdo ima pravico seznaniti se z zbranimi osebnimi podatki, ki se nanašajo nanj, in pravico do sodnega varstva ob njihovi zlorabi.

Komentar k 37. in 38. členu Ustave Republike Slovenije. Ustavne pravice pripadajo vsem državljanom Slovenije ne glede na določbe zakonov in predpisov.

37. člen Ustave Republike Slovenije določa, da je vsebina spletnih komunikacij tajna in da je za kakršnokoli poseganje v to tajnost potrebna odločba sodišča, s čimer so državljani zaščiteni pred nezakonitim prisluškovanjem. Sodna določba mora biti podkrepljena z utemeljitvijo, ki upravičuje njeno nujnost, saj odločba v primeru, da je možno želene podatke zbrati brez poseganja v zasebnost, ne bi smela biti izdana. Odločbe imajo vnaprej določeno obdobje, kar onemogoča dolgotrajne posege v tajnost komunikacij.

38. člen Ustave dopušča javno in tajno zbiranje informacij o spletnih uporabnikih, ne glede na to pa ima vsak državljan pravico do teh informacij.

Ustava določa relativno visoko stopnjo zaščite spletnih uporabnikov, ki pa se je ne spoštuje vedno. Prvi tak primer je že trenutno veljavni Zakon o elektronskih komunikacijah (ZEKom-1A).

Mednarodni trgovski sporazumi

Sporazum ACTA (Anti-Counterfeiting Trade Agreement) je namenjen predvsem zajezitvi kršenja avtorskih pravic. Z njegovim sprejemom v EU bi prišlo do velikih sprememb na več področjih. Med drugim je predvideval boj proti generičnim zdravilom,

patente na hrano in visoko stopnjo regulacije interneta. Njegova vsebina in pogajanja o njem so bila tajna vse do razkritja na strani Wikileaks leta 2008. Po mnenju mnogih domačih [48] in mednarodnih organizacij [49] predstavlja sporazum ACTA grožnjo pravicam in svoboščinam državljanov EU, predvsem domnevi nedolžnosti, svobodi govora in pravici do komunikacijske in informacijske zasebnosti.

Pristop k sporazumu ACTA je po navodilih vlade RS podpisala slovenska veleposlanica v Tokiu Helena Drnovšek Zorko, ki pa se je kasneje javnosti za to opravičila [50]. Sporazum je podpisalo še nekaj članic EU, nadaljnje podpise in ratifikacijo pa je ustavil val protestov po Evropi. Javno ogorčenje in protesti so sprožili odstope držav od sporazuma, v juliju 2012 pa ga je zavrnil tudi Evropski parlament.

Čezatlantski trgovinski sporazum TTIP (Transatlantic Trade and Investment Partnership) je poskus združitve trgov ZDA in EU. Združitev bi lahko negativno vplivala predvsem na državljane EU, saj bi izenačitev predpisov in standardov najverjetneje pomenila nižje okoljske standarde, manj socialnih pravic za evropske delavce in manjšo zaščito zasebnosti evropskih spletnih uporabnikov. Problematična so tudi določila sporazuma, ki ga postavljajo nad državne zakone. Države, ki bi z "neugodno" zakonodajo poskusile ščititi nacionalne posebnosti, oz. interese, bi multinacionalke lahko tožile ne le za predvideni izgubljeni dobiček, temveč celo za izgubljen prihodnji dobiček [51]. Nasprotniki sporazuma opozarjajo na izgubo suverenosti držav, ki tako ne bi mogle preprečiti možnih daljnoročnih posledic npr. uvoza/uporabe gensko spremenjenih živil ali uporabe metode hidravličnega lomljenja (angl. fracking) za pridobivanje zemeljskega plina. Pogajanja o končni vsebini sporazuma med ZDA in EU so tajna in trenutno še tečejo, prav lahko pa bi padla v vodo zaradi razkritja ameriškega vohunjenja po Evropi [52].

Ustavnopravni vidik Ustava RS ne preprečuje ratifikacija mednarodne pogodbe, ki bi bila v nasprotju z Ustavo. Ustavno sodišče ima pravico do presoje ustavnosti mednarodnih pogodb pred njeno ratifikacijo, vendar pa lahko pride do presoje le na predlog predsednika republike, vlade ali tretjine poslancev državnega zbora. Tako je v teoriji lahko ratificirana mednarodna pogodba tudi v nasprotju z Ustavo RS, država pa jo je kljub temu dolžna izpolnjevati, sicer stori mednarodni delikt [47].

Zakon o elektronskih komunikacijah (ZEKom-1A)

Zakon o elektronskih komunikacijah [3], ki je začel veljati v začetku leta 2013, je korenito posegel v ustaljene metode za sledenje spletnim uporabnikom. Po eni strani je s 157. členom zelo omejil uporabo piškotkov, po drugi pa je v 13. poglavju določil, da morajo ponudniki spletnih storitev na lastne stroške, za potrebe Policije in drugih varnostnih organov, hraniti do dve leti stare podatke o prometu, lokacijah in druge povezane podatke. Ustavno sodišče je julija 2014 v sodbi ugotovilo, da to početje ni v skladu z Ustavo in razveljavilo celotno 13. poglavje zakona.

ZEKom-1A - 157. člen (piškotki)

> Shranjevanje podatkov ali pridobivanje dostopa do podatkov, shranjenih v terminalski opremi naročnika ali uporabnika, je dovoljeno samo pod pogojem, da je naročnik ali uporabnik v to privolil po tem, ko je bil predhodno jasno in izčrpno obveščen o upravljavcu in namenih obdelave teh podatkov v skladu z zakonom, ki ureja varstvo osebnih podatkov.
>
> Ne glede na določbe prejšnjega odstavka je dovoljeno tehnično shranjevanje podatkov ali dostop do njih izključno zaradi prenosa sporočila po elektronskem komunikacijskem omrežju ali če je to nujno potrebno za zagotovitev storitve informacijske družbe, ki jo naročnik ali uporabnik izrecno zahtevata.
>
> Če je tehnično izvedljivo in učinkovito ter v skladu z zakonom, ki ureja varstvo osebnih podatkov, se šteje, da uporabnik lahko izrazi svojo privolitev iz prvega odstavka tega člena tudi z uporabo ustreznih nastavitev v brskalniku ali drugih aplikacijah. Privolitev uporabnika ali naročnika pomeni osebno privolitev v skladu z zakonom, ki ureja varstvo osebnih podatkov.
>
> Kadar gre za kršitev pravil o obveščanju in privolitvi posameznika iz prvega odstavka tega člena ter hkrati za kršitev zakona, ki ureja varstvo osebnih podatkov, se uporabljajo določbe tega zakona.
>
> Inšpekcijski nadzor nad izvajanjem določb tega člena opravlja informacijski pooblaščenec.

Komentar k 157. členu ZEKom-1A. Omejitev uporabe piškotkov je naletela na odobravanje spletnih uporabnikov, po drugi strani pa je povzročila težave oglaševalskim ponudnikom, ki so lahko z njihovo uporabo sledili spletnim uporabnikom preko več

spletnih strani. Pred veljavnostjo tega zakona je bilo to početje zakonito in zelo razširjeno, spletni uporabniki pa so se poskušali sledenju izogniti z brisanjem piškotkov v brskalnikih ali z uporabo t.i. *zasebnostnega načina* delovanja brskalnika.

Novi zakon določa, da lahko ponudniki oglaševanja uporabljajo spletne piškotke le ob predhodnem izrecnem soglasju uporabnika. Ta določba je povzročila veliko bojazni, saj se je predvidevalo, da veliko uporabnikov ne bo želelo sodelovati z oglaševalci in da se bo veliko privolitev v sledenje zaradi uporabe zasebnega načina delovanja brskalnika izbrisalo. S tem bi se velikemu številu slovenskih spletnih strani občutno znižal prihodek od oglaševanja in zmanjšala učinkovitost porabe denarja v oglaševanju. To bi dalo tujim, predvsem ameriškim, podjetjem, konkurenčno prednost. Dobro leto po začetku veljavnosti tega zakona se zdi, da so bile te bojazni večinoma pretirane.

Informacijski pooblaščenec

Informacijski pooblaščenec je samostojen in neodvisen državni organ, ki je med drugim pristojen za varstvo osebnih podatkov. Od njegove ustanovitve leta 2003 naprej ima ta organ vedno več pristojnosti in pooblastil na področjih informacij javnega značaja in varstva osebnih podatkov - med drugim izvaja inšpekcijski nadzor nad 157. členom ZEKom-1A, ki ureja uporabo piškotkov in podobnih tehnologij na spletnih straneh.

A.3.2 Zasebnost in oglaševanje

Ciljano oglaševanje, ki omogoča bolj ekonomično porabo denarja v spletnem oglaševanju, potrebuje za svoje delovanje informacije o populaciji spletnih uporabnikov. Za zbiranje informacij o uporabnikih se najpogosteje uporabljajo neinvazivne metode, ki temeljijo na spremljanju, analizi in agregaciji preteklih aktivnosti posameznih spletnih uporabnikov, čemur pravimo tudi profiliranje.

V socialnih omrežjih lahko ciljno populacijo za oglaševalsko akcijo definiramo kot poimenski seznam uporabnikov ali pa kot množico uporabnikov, ki ustrezajo določenim kriterijem, kot so na primer spol, starostni okvir in povezave s stranmi filmov, blagovnih znamk itd. Raba uporabniških podatkov za potrebe oglaševanja je navadno omenjena v pogojih uporabe socialnih omrežij, tako da se v tem primeru uporabniki sami odrečejo pravici do zasebnosti.

Spletne oglaševalske mreže, ki navadno vključujejo večje število spletnih strani in oglaševalcev, navadno ne razpolagajo s tako podrobnimi podatki o spletnih uporabnikih. Obiskovalci spletnih strani so s stališča mreže anonimni, saj je o vsakem znana le

identifikacijska številka in podatki o obiskih spletnih strani. Kljub temu, da so spletni uporabniki v oglaševalski mreži anonimni, pa se sledenje uporabniku brez njegovega privoljenja obravnava kot kršitev zasebnosti.

Dodatek: Analiza vpliva postavitve oglasov na spletni strani na njihovo učinkovitost pri ciljanem oglaševanju

B

B.1 Učinkovitost oglasov glede na njihovo pozicijo na spletnih straneh

B.1.1 Študije uporabnosti spletnih strani

Napredne študije uporabnosti spletnih strani temeljijo na podlagi podatkov o gibanju oči obiskovalcev. Za sledenje očem se uporabljajo spletne kamere, ki so pritrjene na posebno vrsto čelade ali zaslon računalnika. Poskusi navadno zajemajo manjše število (do nekaj deset) uporabnikov, zbrani podatki pa natančno opisujejo, kam in koliko časa je uporabnik gledal.

Testni subjekti so navadno postavljeni pred nalogo:

- čim hitreje prepoznati istovetnost spletne strani ali
- poiskati ceno določenega izdelka na spletni strani.

Na podlagi hitrosti in učinkovitosti uporabnikov pri reševanju nalog načrtovalci določijo optimalen izgled spletne strani (t.j. postavitev strani ter izbira ustreznih grafičnih in tekstovnih elementov) [53].

Velik del raziskav na tem področju se ukvarja z optimizacijo spletnih iskalnikov. Poleg preprostega in močnega uporabniškega vmesnika je pri spletnem iskalniku najpomembnejši prikaz rezultatov (angl. search engine result page - SERP). Pri analizah učinkovitosti SERP so zato testni uporabniki največkrat postavljeni pred nalogo čimprej najti informacije o določeni entiteti (npr. lokacija znanega prodajalca avtomobilov) [54].

Velika količina informacij nas sili v vedno hitrejše in bolj površno branje besedil na spletu. Duggan in Payne [55] sta s sledenjem očem bralcev znanstvenih člankov potrdila svojo teorijo, da ljudje navadno beremo besedilo vse dokler je njegova informativna vrednost visoka, nakar skočimo na naslednji odstavek.

Pri iskanju informacij se dandanes vedno bolj zanašamo na spletne iskalnike. Pan in sod. [56] so pokazali, da ima velik vpliv na našo uspešnost pri iskanju informacij prav vrstni red rezultatov. V poskusih so s ponovnim rangiranjem rezultatov iskalnika Google umetno znižali njihovo kvaliteto. Testni uporabniki so sicer opazili, da najbolje rangirani rezultati niso najboljši, vendar se pri reševanju nalog niso obnesli enako dobro, kot pri originalni razvrstitvi rezultatov iskanj.

Sherman [57] je na podlagi sledenja očem uporabnikov spletnega iskalnika Google izrisal najpogostje obiskane dele strani z rezultati in pri tem prvi opazil dva močna

vodoravna trakova in enega navpičnega, vsi skupaj pa sestavljajo vzorec, podoben črki *F*. Podobne vzorce so na tej in na drugih spletnih straneh kasneje našli tudi drugi raziskovalci [58].

Cutrell in Guan [59] sta ugotovila, da lahko na uporabnost spletnega iskalnika pomembno vplivajo tudi podrobnosti, kot je na primer velikost odlomka besedila ob vsakem rezultatu. V primeru, da uporabnik/ca prek poizvedbe išče informacije, so dolgi odlomki dobrodošli, če pa uporablja spletni iskalnik le za navigacijo do želene spletne strani, pa je bolje, da so ti odlomki čimkrajši.

Pomembnost razlikovanja med različnimi cilji uporabnikov so pokazali tudi Buscher in sod. [60]. Pri iskanju informacij so se uporabniki spletnih strani bolj osredotočali na sredino spletne strani, kjer se nahaja vsebina. Za razpoznavanje spletne strani pa so obiskovalci bolj uporabljali informacije, ki so se nahajale zgoraj in levo zgoraj.

Pri analizah obiskovalcev spletnih strani se je med drugim izkazalo še, da:

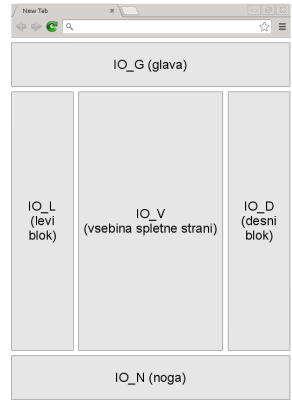
- k učinkovitejšemu iskanju informacij veliko prispeva uporaba podnaslovov in naštevanje v alineah,
- manjša velikost pisave in krajši odstavki spodbujajo obiskovalce/ke k intenzivnejšemu branju vsebine in
- oglasi, ki so vizualno drugačni in/ali daleč od vsebine spletne strani, ostanejo ponavadi neopaženi.

Zaradi poplave oglasov na svetovnem spletu je pri spletnih uporabnikih prišlo do dveh zanimivih pojavov:

- oglasna slepota (angl. banner blindness) - Uporabniki namerno ne gledajo v dele spletnih strani, ki spominjajo na oglase [61].
- izogibanje animacijam (angl. animation avoidance) - Uporabniki ignorirajo svetleče, migetajoče in utripajoče elemente na spletni strani.

B.1.2 Učinkovitost oglasov glede na pozicijo na spletnih straneh

Odločitev, kje na spletni strani bodo umeščeni oglasi, je popolnoma v rokah založnika. Kljub temu pa je koristno vedeti, ali in kako lahko lokacija oglasa na spletni strani vpliva na njegovo učinkovitost.



Slika B.1

Prikaz razdelitve tipične spletne strani na 5 interesnih območij (IO): glava, levi blok, vsebina, desni blok in noga.

Podatki o pozicijah oglasov in klikih na njih

Podatke o pozicijah oglasnih blokov na spletnih straneh smo pridobili iz arhiva spletnih strani[1].

Pri primerjavi oglaševalskih pozicij smo se oprli na standardno metodologijo, ki se uporablja za ocenjevanje uporabnosti spletnih strani in njihovo optimizacijo. Tipično spletno stran smo razdelili na interesna območja IO [53]. Razdelitev tipične spletne strani na interesna območja je prikazana na sliki B.1. IO si v izvorni kodi HTML sledijo v naslednjem zaporedju:

1. IO_G (glava),
2. IO_L (levi blok),
3. IO_V (vsebina spletne strani),
4. IO_D (desni blok) in
5. IO_N (noga).

[1]Internet Archive Wayback Machine https://archive.org/

Potrebno je opozoriti, da na dosti spletnih straneh niso prisotna vsa interesna območja. Največkrat manjkata na spletni strani levi (IO_L) ali desni blok (IO_D), kar pa ne vpliva na našo študijo.

Kot vhodne podatke za medsebojno primerjavo IO spletne strani smo namesto podatkov o gibanju oči spletnih uporabnikov uporabili podatke o prikazih in klikih na oglase iz enakega časovnega obdobja, kot je bilo uporabljeno v poglavju 5 za evalvacijo sistemov za priporočanje. V primerjavo dveh IO smo vključili vse klike, za katere velja:

- v vsakem od IO je umeščen najmanj en oglasni blok in
- klik na oglas se je zgodil na enem izmed primerjanih IO.

Tema dvema pogojema je ustrezalo le okoli 20% dogodkov (ogledov spletnih strani z oglasi), zaradi česar smo v primerjavi s podatki, uporabljenimi v poglavju 5, v analizo vključili dogodke 4x večjega števila uporabnikov.

V testnem obdobju se je za izbiro oglasov za prikaz na spletnih straneh uporabljalo kontekstualno ciljanje. To pomeni, da so se spletnim uporabnikom prikazovali oglasi, ki so se vsebinsko v čimvečji meri ujemali z vsebino obiskane spletne strani.

Rezultati

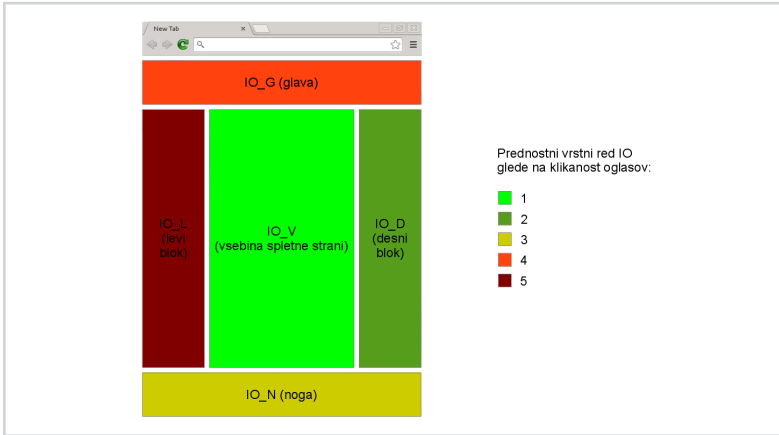
V tabeli B.1 je prikazana medsebojna primerjava interesnih območij spletne strani. Hiter pregled rezultatov pokaže, da lahko glede na dobljene rezultate interesna območja razvrstimo po klikanosti oglasov od najbolj primernega IO (IO_V) do najmanj primernega (IO_L). Razvrstitev interesnih območij po klikanosti je prikazana na sliki B.2 in je sledeča:

1. IO_V (vsebina spletne strani),
2. IO_D (desni blok),
3. IO_N (noga),
4. IO_G (glava) in
5. IO_L (levi blok).

Tabela B.1

Medsebojna primerjava interesnih območij spletne strani glede na klikanost oglasov. V primerjavi dveh IO so zajeti le kliki na spletnih straneh, na katerih so se pojavili oglasi v obeh IO, klik pa se zgodil v enem izmed njiju.

klikov skupaj	IO 1	klikov v IO 1	IO 2	klikov v IO 2
138	IO_G	*136*	IO_L	2
8	IO_G	1	IO_V	*7*
2116	IO_G	276	IO_D	*1840*
57	IO_G	24	IO_N	*33*
128	IO_L	62	IO_V	*66*
440	IO_L	5	IO_D	*435*
229	IO_L	53	IO_N	*176*
556	IO_V	*345*	IO_D	211
344	IO_V	*231*	IO_N	113
1900	IO_D	*1629*	IO_N	271



Slika B.2

Prikaz razdelitve tipične spletne strani na 5 interesnih območij (IO): glava, levi blok, vsebina, desni blok in noga. Z barvami je predstavljen prednostni vrstni red IO glede na njihovo uspešnost po klikanosti oglasov.

Zanimiva je še primerjava dveh združenih interesnih območij ZIO (tabela B.2), ki združujeta IO glede na njihov vrstni red v izvorni kodi HTML spletne strani:

- ZIO 1: glava in levi blok spletne strani,
- ZIO 2: vsebina, desni blok in noga spletne strani.

Tabela B.2

Medsebojna primerjava združenih interesnih območij ZIO. V prvo združeno interesno območje sta združena glava IO_G in levi blok spletne strani IO_L, v drugo pa vsebina IO_V, desni blok IO_D in noga IO_N.

klikov skupaj	ZIO 1	klikov v ZIO 1	ZIO 2	klikov v ZIO 2
2523	IO_G IO_L	351	IO_V IO_D IO_N	*2172*

Oglasi, umeščeni v združeno interesno območje ZIO 1, so bili kliknjeni le v 14% primerih, oglasi v ZIO 2 pa v 86% primerih.

Ugotovitve

Naši rezultati se deloma skladajo z ugotovitvami študij o uporabnosti spletnih strani, ki temeljijo na sledenju očem spletnih uporabnikov.

Ugotovili smo, da so oglasi, umeščeni v samo vsebino spletne strani (interesno območje IO_V), največkrat kliknjeni, kar sovpada z gibanjem oči uporabnikov, ki iščejo informacije na spletnih straneh [60].

Pri večini študij, ki je med seboj primerjala levi in desni blok spletne strani, so raziskovalci prišli do zaključka, da uhajajo pogledi uporabnikov in uporabnic večkrat na levo stran [62] kot na desno. Naši rezultati (tabela B.1) kažejo nasprotno sliko, in sicer, da so oglasi na desni kliknjeni večkrat, kot na levi. Razlogov za to je več. Veliko je odvisno od postavitve posameznih spletnih strani, tekstovnih in grafičnih elementov, barvnih kombinacij itd, največji pa je verjetno prevladujoči delež jezikov, ki se jih bere z leve proti desni.

Poleg same spletne strani pa je veliko odvisno tudi od namenov uporabnika/ce. Spletni uporabniki so ciljno usmerjeni [63]: nekateri iščejo informacije (npr. naslov podjetja ali cena izdelka), drugi pa le navigirajo v iskanju prave spletne strani. To se odraža tudi v njihovem obnašanju oz. pregledovanju obiskane spletne strani [60]. Pri

iskanju informacij so raziskovalci opazili zelo značilno pregledovanje spletne strani z gibanjem pogleda v obliki črke *F* [57, 58], kar bi lahko razložilo, zakaj sta se v naših poskusih interesni območji IO_G (glava) in IO_L (levi blok) izkazala za najmanj klikana (tabela B.2). S prvim pogledom uporabniki navadno poiščejo naslov ali večji kos besedila, kar ustreza levemu zgornjemu kotu črke *F*, vsi nadaljnji pogledi uporabnikov pa so usmerjeni v desno smer in/ali dol. Ob takem gibanju ne bi pogled uporabnika/ce nikoli niti prišel do interesnih območij IO_G in IO_L.

Ta teorija bi lahko do določene mere pojasnila nenavadno slabe izmerjene kvalitete "najboljših" priporočil sistemov za priporočanje z vsebinskim izbiranjem, opisane v tabelah 5.5-5.7 v razdelku 5.4.2. Predpostavljamo lahko namreč, da se profili spletnih uporabnikov v veliki meri ujemajo s kontekstualno vsebino spletnih strani. Oglasi, ki so se prikazovali na obiskanih spletnih straneh, so bili ciljani kontekstualno, in sicer tako, da so se v prvem oglasnem bloku prikazali "najboljši" kontekstualni oglasi, v drugem malo slabši itn. Vrstni red oglasnih blokov na oglasih je definiran z njihovim vrstnim redom v izvorni kodi spletne strani (glej razdelek B.1.2).

To pomeni, da so se na vrhu in v levih blokih prikazovali oglasi, ki so se najbolje ujemali z vsebinami spletnih strani in posledično tudi s profili spletnih uporabnikov. Ob predpostavki, da veliko spletnih uporabnikov pregleduje spletne strani z gibanjem pogleda v obliki črke *F*, pa bi bili ravno ti oglasi tudi največkrat spregledani in zato manj pogosto kliknjeni.

Zelo naivno bi bilo predvidevati, da uporabljajo vsi spletni uporabniki enako strategijo pri pregledovanju vsebin spletnih strani, zato predpostavljamo, da obstajajo za nenavadne rezultate priporočil omenjenih sistemov za priporočanje tudi drugi razlogi.

Dodatek: Profiliranje z uporabo sinusne regresije

C

C.1 Profilirni algoritmi na osnovi sinusne regresije

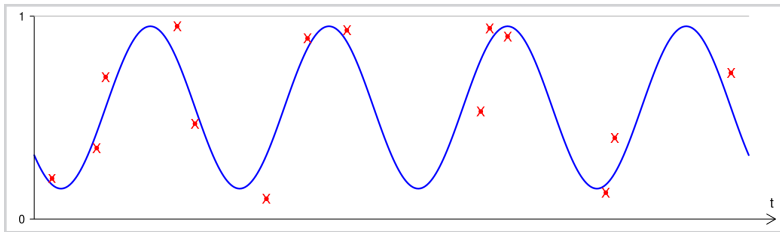
C.1.1 Aproksimacija gibanja ocen konceptov s sinusno regresijo

Sinusna regresija

Gibanje ocen posameznih konceptov smo poskušali aproksimirati tudi s sinusno regresijo (glej sliko C.1), ki se uporablja predvsem za modeliranje periodičnih gibanj vrednosti. Uporabo sinusne regresije za modeliranje interesov spletnih uporabnikov je vodila intuicija, da se obnašanje uporabnikov čez dan, ko so v službi ali šoli, bistveno razlikuje od večernih aktivnosti, ki so večinoma namenjene sprostitvi in zabavi. Po drugi strani pa smo pričakovali določeno mero podobnosti med uporabnikovimi aktivnostmi ob istih urah v dnevu.

Slika C.1

Prikaz delovanja sinusne regresije. Z rdečo so označene ocene izbranega koncepta v dogodkih iz učnega klikotoka. Gibanje ocen je aproksimirano s sinusoido, ki se uporablja kot napovedni model za vrednosti ocen koncepta pri evalvaciji uporabnikovega profila.



Osnovni algoritem, ki gradi uporabniške profile s sinusno regresijo, optimizira z minimizacijo kvadratne napake 4 parametre v enačbi C.1:

- a - amplituda gibanja ocen,
- b - frekvenca sinusoide,
- c - horizontalni ali fazni zamik sinusoide in
- d - vertikalni zamik.

$$y = a * cos(b * x + c) + d \tag{C.1}$$

Pri analizi zgrajenih profilov se je pokazalo, da velika večina frekvenc v modelih približno ustreza dnevni $b \approx \frac{1}{1dan}$ ali tedenski periodi $b \approx \frac{1}{1teden}$. To je vodilo k razvoju več različic tega profilirnega algoritma.

Sinusna regresija s fiksno periodo

Uporaba fiksne periode v postopku sinusne regresije občutno olajša in pohitri optimizacijo ostalih parametrov. Razvili smo več različic profilirnih algoritmov, ki aproksimirajo gibanje ocen konceptov:

- s sinusoido s *fiksno dnevno periodo*. Tu je bila glavna motivacija razlikovanje med uporabnikovimi interesi ob različnih urah v dnevu in pojavljanje podobnih aktivnosti iz dneva v dan ob istih urah,
- s sinusoido s *fiksno tedensko periodo*. Z uporabo tedenske periode smo hoteli predvsem izkoristiti razliko v aktivnostih uporabnika med delavniki in med vikendom,
- z *izbiro najprimernejše fiksne periode*, ki razlikuje med vsakodnevnimi in tedenskimi aktivnostmi spletnih uporabnikov. Gibanje ocen vsakega koncepta smo skušali aproksimirati s sinusoido z dnevno in s sinusoido s tedensko periodo. Aproksimacijo z večjo vsoto kvadratnih napak smo zavrgli, drugo pa smo dodali v uporabnikov profil in jo uporabljali za napovedovanje njegovih interesov.

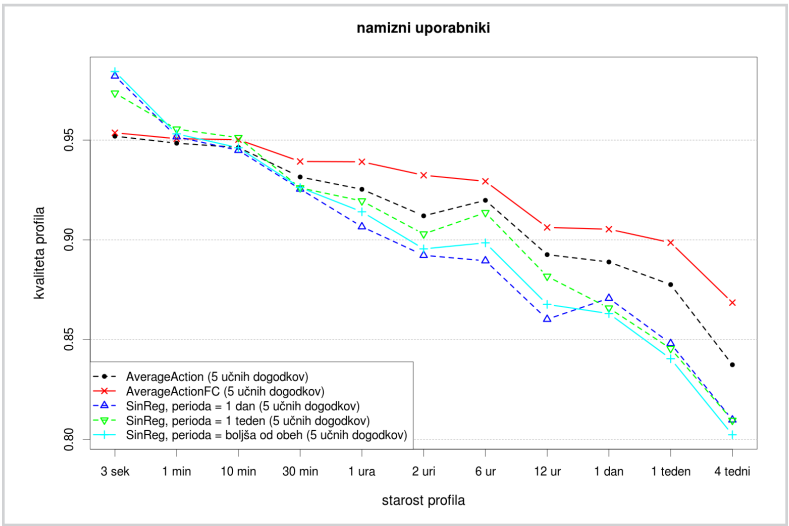
C.1.2 Kvaliteta profilov na osnovi sinusne regresije

Gradnja profilov z uporabo sinusne regresije (glej razdelek C.1.1) se ni izkazala za uspešno. Kot je razvidno iz slik C.2 - C.5, bi lahko bil tak pristop koristen za napovedovanje uporabnikovih kratkoročnih interesov, pri napovedovanju srednje- in dolgoročnih interesov pa je slabši tudi od metode AverageAction, ki ne uporablja niti časovnega pozabljanja, niti popravljanja profilov s prototipi.

Izkazalo se je, da lahko s sinusno regresijo zelo dobro modeliramo uporabnikove kratkoročne interese - celo nekoliko bolje kot z algoritmom LastAction, ki se je sicer pri tem izkazal za najprimernejšega. V medsebojni primerjavi vseh različic se je pokazalo, da sta sinusni regresiji z uporabo fiksne dnevne in tedenske perioda približno enako primerni za modeliranje uporabnikovih interesov, dvojna sinusna regresija pa daje občutno slabše rezultate.

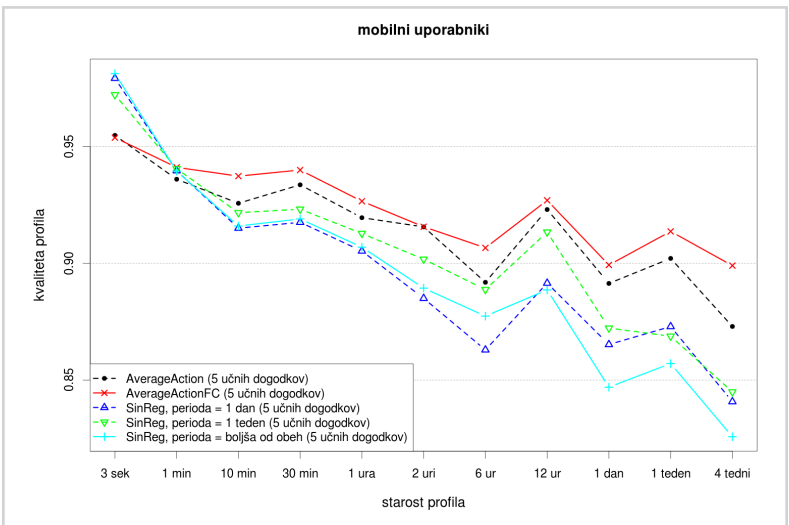
Pri modeliranju srednjeročnih in dolgoročnih interesov spletnih uporabnikov se je sinusna regresija izkazala za nekoristno, saj je dajala celo slabše rezultate od metode AverageAction, ki pri gradnji profila le povpreči ocene konceptov preko vseh dogodkov

iz učnega klikotoka. Tudi uporaba časovnega pozabljanja ni izboljšala kvalitet srednje- in dolgoročnih napovedi.



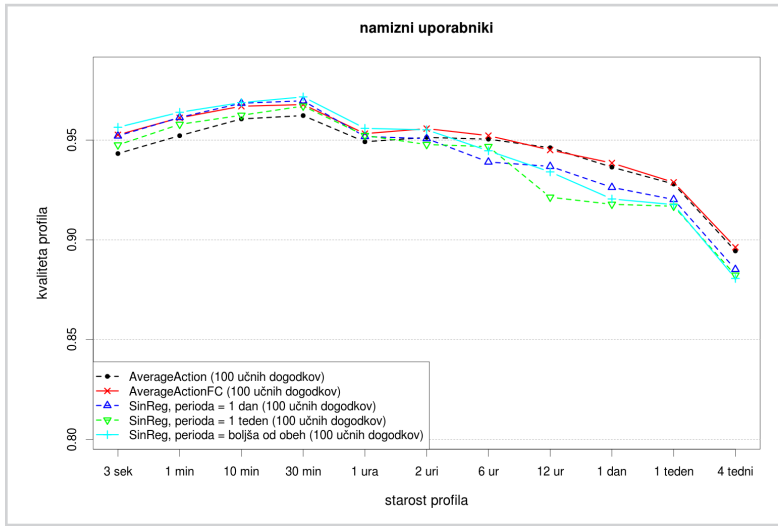
Slika C.2

Primerjava metode AverageActionFC s profiliranjem z uporabo sinusne regresije na populaciji namiznih uporabnikov. Uporabljeni so kratki učni klikotoki dolžine $n = 5$.



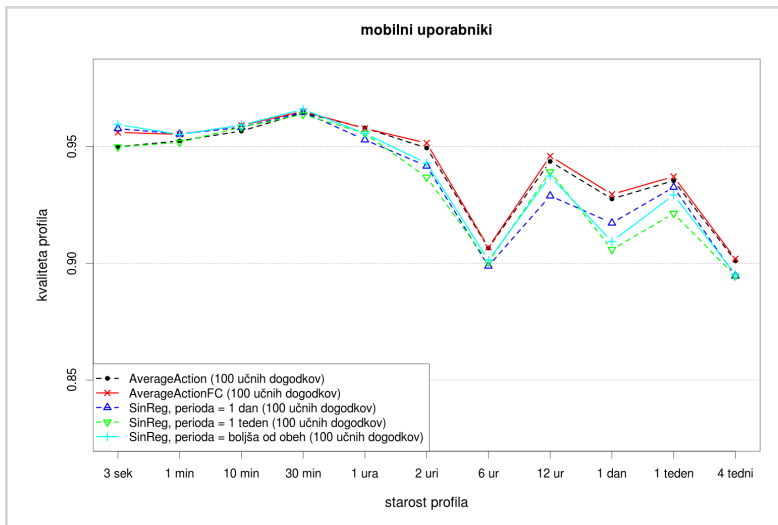
Slika C.3

Primerjava metode AverageActionFC s profiliranjem z uporabo sinusne regresije na populaciji mobilnih uporabnikov. Uporabljeni so kratki učni klikotoki dolžine $n = 5$.



Slika C.4

Primerjava metode AverageActionFC s profiliranjem z uporabo sinusne regresije na populaciji namiznih uporabnikov. Uporabljeni so dolgi učni klikotoki dolžine $n = 100$.



Slika C.5

Primerjava metode AverageActionFC s profiliranjem z uporabo sinusne regresije na populaciji mobilnih uporabnikov. Uporabljeni so dolgi učni klikotoki dolžine $n = 100$.

LITERATURA

[1] Jef I. Richards and Catharine M. Curran. Oracles on "advertising": Searching for a definition. *Journal of Advertising*, 31(2):63–77, 2002. doi: 10.1080/00913367.2002.10673667. URL http://www.tandfonline.com/doi/abs/10.1080/00913367.2002.10673667.

[2] Zlatko Jančič, Vesna Žabkar, Miro Kline, Klement Podnar, Tanja Kamin, Domen Bajde, Dejan Verčič, and Urša Golob. *Oglaševanje*. Založba FDV, 2013. URL http://knjigarna.fdv.si/knjige/komunikologija/moj-zbirka-za-marketing-in-odnose-z-javnostmi/i_599_oglasevanje.

[3] Državni zbor Republike Slovenije. Zakona o elektronskih komunikacijah (zekom-1). Uradni list Republike Slovenije, December 2012. URL http://www.uradni-list.si/1/content?id=111442. (23.07.2014).

[4] Samy Kamkar. evercookie, 2010. URL http://samy.pl/evercookie/. (17.03.2013).

[5] Nataša Pirc Musar. Informacijski pooblaščenec izdal smernice glede uporabe piškotkov. IP-RS, March 2013. URL http://www.ip-rs.si/novice/detajl/informacijski-pooblascenec-izdal-smernice-glede-uporabe-piskotkov/. (15.10.2014).

[6] Daniel Billsus and Michael J. Pazzani. User modeling for adaptive news access. *User Modeling and User-Adapted Interaction*, 10(2-3):147–180, June 2000. doi: 10.1023/A:1026501525781. URL http://link.springer.com/article/10.1023/A%3A1026501525781.

[7] Elnaz Davoodi, Keivan Kianmehr, , and Mohsen Afsharchi. A semantic social network-based expert recommender system. *Applied intelligence*, 39(1):1–13, July 2013. doi: 10.1007/s10489-012-0389-1. URL http://link.springer.com/article/10.1007/s10489-012-0389-1.

[8] Ahu Sieg, Bamshad Mobasher, and Robin Burke. Learning ontology-based user profiles: A semantic approach to personalized web search. *IEEE Intelligent Informatics Bulletin*, 8(1):7–18, November 2007.

[9] Yiouli Kritikou, Panagiotis Demestichas, Evgenia Adamopoulou, Konstantinos Demestichas, Michael Theologou, and Maria Paradia. User profile modeling in the context of web-based learning management systems. *Journal of Network and Computer Applications*, 31(4):603–627, November 2008. doi: 10.1016/j.jnca.2007.11.006. URL http://www.sciencedirect.com/science/article/pii/S1084804507000720.

[10] Daniela Petrelli, Antonella De Angeli, and Gregorio Convertino. A user-centered approach to user modeling. In *UM99 User Modeling - Proceedings of the Seventh International Conference*, pages 255–264. Springer Vienna, 1999. doi: 10.1007/978-3-7091-2490-1_25. URL http://link.springer.com/chapter/10.1007/978-3-7091-2490-1_25.

[11] Reynol Junco. Comparing actual and self-reported measures of facebook use. *Computers in Human Behavior*, 29(3):626–631, May 2013. doi: 10.1016/j.chb.2012.11.007. URL http://www.sciencedirect.com/science/article/pii/S0747563212003123.

[12] Edward Thomas, Jeff Z. Pan, Stuart Taylor, Yuan Ren, Nophadol Jekjantuk, and Yuting Zhao. Semantic advertising for web 3.0. In Tanja Zseby, Reijo Savola, and Marco Pistore, editors, *Future Internet - FIS 2009*, volume 6152 of *Lecture Notes in Computer Science*, pages 96–105. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-14955-9. doi: 10.1007/978-3-642-14956-6_9. URL http://dx.doi.org/10.1007/978-3-642-14956-6_9.

[13] Adriana S. Vivacqua, Jonice Oliveira, and Jano M. de Souza. i-prose: Inferring user profiles in a scientific context. *The Computer Journal*, 52(7):789–798, October 2009. doi: 10.1093/comjnl/bxp002. URL

http://comjnl.oxfordjournals.org/content/52/7/789.abstract.

[14] The open directory project, 1998. URL http://www.dmoz.org/. (03.08.2012).

[15] Victoria Eyharabide and Analía Amandi. Ontology-based user profile learning. *Applied Intelligence*, 36 (4):857–869, June 2012. doi: 10.1007/s10489-011-0301-4. URL http://link.springer.com/article/10.1007/s10489-011-0301-4.

[16] Stuart E. Middleton, Nigel R. Shadbolt, and David C. De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):54–88, January 2004. doi: 10.1145/963770.963773. URL http://dl.acm.org/citation.cfm?id=963773.

[17] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 321–328. ACM, July 2004.

[18] Mariam Daoud, Lynda-Tamine Lechani, and Mohand Boughanem. Towards a graph-based user profile modeling for a session-based personalized search. *Knowledge and Information Systems*, 21(3):365–398, December 2009. doi: 10.1007/s10115-009-0232-0. URL http://link.springer.com/article/10.1007/s10115-009-0232-0.

[19] James Allan. Hard track overview in trec 2003 high accuracy retrieval from documents. In E. M. Voorhees and Lori P. Buckland, editors, *The Twelfth Text Retrieval Conference (TREC 2003)*, pages 24–37, 2004.

[20] Daniela Godoy and Analía Amandi. Modeling user interests by conceptual clustering. *Information Systems*, 31(4-5):247–265, June-July 2006. doi: 10.1016/j.is.2005.02.008. URL http://www.sciencedirect.com/science/article/pii/S0306437905000335.

[21] Daniela Godoy and Analía Amandi. Interest drifts in user profiling: A relevance-based approach and analysis of scenarios. *The Computer Journal*, 52(7):771–788, October 2009. doi: 10.1093/comjnl/bxm107. URL http://comjnl.oxfordjournals.org/content/52/7/771.abstract.

[22] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967. URL http://projecteuclid.org/euclid.bsmsp/1200512992.

[23] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[24] Prasanna Ganesan, Hector Garcia-Molina, and Jennifer Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems (TOIS)*, 21(1):64–93, January 2003. doi: 10.1145/635484.635487. URL http://dl.acm.org/citation.cfm?id=635487.

[25] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945. doi: 10.2307/3001968.

[26] Luca Passani. Wurfl - wireless universal resource file, 2001. URL http://wurfl.sourceforge.net/. (11.09.2011).

[27] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[28] Gediminas Adomavicius and Alexander Tuzhilin. Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *xxx*, 6:73–96, 2005. doi: 10.4000/revus.1841. URL http://revus.revues.org/1841.

[29] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46. ACM, September 2010.

[30] M. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27:313–331, 1997. doi: xxx. URL http://revus.revues.org/1841.

[31] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. IEEE, 2008.

[32] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009. doi: 10.1109/MC.2009.263. URL http://www.computer.org/csdl/mags/co/2009/08/mco2009080030-abs.html.

[33] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. Sorec: social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 931–940. ACM, 2008.

[34] Seung-Taek Park, David Pennock, Omid Madani, Nathan Good, and Dennis DeCoste. Naïve filterbots for robust cold-start recommendations. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 699–705. ACM, 2006.

[35] Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Steffen Rendle, and Lars Schmidt-Thieme. Learning attribute-to-feature mappings for cold-start recommendations. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 176–185. IEEE, 2010.

[36] An-Te Nguyen, Nathalie Denos, and Catherine Berrut. Improving new user recommendations with rule-based induction on cold user data. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 121–128. ACM, 2007.

[37] Matthias Braunhofer, Marius Kaminskas, and Francesco Ricci. Location-aware music recommendation. *International Journal of Multimedia Information Retrieval*, 2(1):31–44, 2013.

[38] Linas Baltrunas, Bernd Ludwig, and Francesco Ricci. Matrix factorization techniques for context aware recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 301–304. ACM, 2011.

[39] Peter Forbes and Mu Zhu. Content-boosted matrix factorization for recommender systems: experiments with recipe recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 261–264. ACM, 2011.

[40] Daniel J Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3):477–564, 2006.

[41] Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. MyMediaLite: A free recommender system library. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys 2011)*, pages 305–308. ACM, 2011.

[42] Rong Pan, Yunhong Zhou, Bin Cao, Nathan Nan Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. One-class collaborative filtering. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 502–511. IEEE, 2008.

[43] Slovenska oglaševalska zbornica. Slovenski oglaševalski kodeks, October 2009. URL http://www.soz.si/uploads/files/SOZ_SOK_SLO.pdf. (07.06.2014).

[44] Vrhovno sodišče Republike Slovenije. Sodba iii ips 121/98. Sodstvo Republike Slovenije, October 1998. URL http://www.sodisce.si/znanje/sodna_praksa/vrhovno_sodisce_rs/31086/. (09.08.2014).

[45] Emilio Serrano, Jose M. Such, Juan A. Botía, and Ana García-Fornes. Strategies for avoiding preference profiling in agent-based e-commerce environments. *Applied Intelligence*, 40(1):127–142, January 2014. doi: 10.1007/s10489-013-0448-2. URL http://link.springer.com/article/10.1007/s10489-013-0448-2.

[46] Ustava republike slovenije, December 1991. URL http://www.us-rs.si/o-sodiscu/pravna-podlaga/ustava/. (23.07.2014).

[47] Mirjam Škrk. Ustavnosodna presoja mednarodnih pogodb. *Revus. Revija za ustavno teorijo in filozofijo prava / Journal for Constitutional Theory and Philosophy of Law*, 6:73–96, 2006. doi: 10.4000/revus.1841. URL http://revus.revues.org/1841.

[48] Nataša Pirc Musar. Informacijski pooblaščenec o acta. IP-RS, February 2012. URL https://www.ip-rs.si/novice/detajl/informacijski-pooblascenec-o-acta/. (22.07.2014).

[49] Acta: A global threat to freedoms (open letter). Free Knowledge Institute, December 2009. URL http://freeknowledge.eu/acta-a-global-threat-to-freedoms-open-letter. (22.07.2014).

[50] Helena Drnovšek Zorko. Zakaj sem podpisala acta-o. Metina lista, January 2012. URL http://metinalista.si/zakaj-sem-podpisala-acta-o/. (22.07.2014).

[51] George Monbiot. This transatlantic trade deal is a full-frontal assault on democracy. The Guardian, November 2013. URL http://www.theguardian.com/commentisfree/2013/nov/04/us-trade-deal-full-frontal-assault-on-democracy. (23.07.2014).

[52] Amitabh Pal. U.s. spying on germany causes crisis. The Progressive, July 2014. URL http://www.progressive.org/news/2014/07/187780/us-spying-germany-causes-crisis. (23.07.2014).

[53] Mark Russell. Using eye-tracking data to understand first impressions of a website. *Usability News*, 7(1): 1–14, 2005.

[54] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, 59(7):1041–1052, 2008.

[55] Geoffrey B Duggan and Stephen J Payne. Skim reading by satisficing: evidence from eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1141–1150. ACM, 2011.

[56] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3): 801–823, 2007.

[57] Chris Sherman. A new f-word for google search results. Search Engine Watch, March 2005. URL http://searchenginewatch.com/sew/news/2066806/a-new-f-word-google-search-results. (13.01.2015).

[58] Jakob Nielsen. F-shaped pattern for reading web content. Nielsen Norman Group, April 2006. URL http://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/. (13.01.2015).

[59] Edward Cutrell and Zhiwei Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 407–416. ACM, 2007.

[60] Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30. ACM, 2009.

[61] Jakob Nielsen. Banner blindness: Old and new findings. Nielsen Norman Group, August 2007. URL http://www.nngroup.com/articles/banner-blindness-old-and-new-findings/. (13.01.2015).

[62] Jakob Nielsen. Horizontal attention leans left. Nielsen Norman Group, April 2010. URL http://www.nngroup.com/articles/horizontal-attention-leans-left/. (13.01.2015).

[63] Jakob Nielsen and Don Norman. Making web advertisements work. Nielsen Norman Group, May 2003. URL http://www.nngroup.com/articles/making-web-advertisements-work/. (13.01.2015).